Vol. IX
Part II

The British Journal of Statistical Psychology

November
1956

'SPEARMAN'S RHO' AND THE MATCHING PROBLEM

By D. E. BARTON and F. N. DAVID
University College, London

Abstract. The following paper describes a possible model, based on a distortion of the random sequence, which may be used as a hypothesis alternative to the hypothesis of randomness. The model is applied to the case of ' Spearman's rho '; but may be adopted as an alternative to the random sequence for any bivariate criterion.

1. The concept of the power function introduced by Neyman and Pearson is one which has proved of increasing value in statistical techniques. In essence the idea is simple: it devolves naturally from the statement that, in any statistical problem, it is necessary to specify not only H_0, the hypothesis tested, but H_1, the class of hypotheses which one would believe to be tenable, if the hypothesis tested were rejected as untenable. According to the Neyman-Pearson school all analysis starts from this point. Once H_0 and H_1 are specified, a criterion which will test H_0 and be sensitive to H_1 can be laid down. The probabilities of rejecting H_0 when it is tenable on the data analysed are the usual familiar significance levels; the probability of rejecting H_0 when the true underlying hypothesis is not H_0 but one of the class of H_1, we 'shall call the power of the test of H_0 with respect to H_1.'

Statistical tests on ranked observations are usually either tests for randomness, or tests of independence, or both. And, since randomness and independence have many facets, it would seem to be important that the alternative hypotheses should be clearly specified. Attempts at this last problem, however, have resulted in the past in power functions of considerable complexity—chiefly, it is suggested, because research workers have endeavoured to assume something about the probability distribution function of the variables before assigning ranks. In the present note we put forward a model for the alternative hypothesis which is based solely on a distortion of the random sequence. For the sake of example we illustrate our procedure on the approximate distribution of what is commonly known as ' Spearman's ρ '.

2. We begin by exploring the connection between Spearman's ρ and the matching problem under the null hypothesis. It is supposed that there are available n pairs of observations $\{x_i y_i\}$. If the ranks of the x's are arranged in ascending numerical order, and if r_i is the y rank corresponding to the ith x rank, we may write our series as $\{i, r_i\}$, $i = 1,2,..., n$. A matched pair will occur when the ranks of x and of y are the same, e.g., when $r_i = i$. Under the null hypothesis that x and y are randomly and independently associated the probability of R matches has been known for some time. Writing this probability as $p_0(R)$ we have

$$p_0(R) = \sum_{i=0}^{n-R} \frac{(-1)^j}{R!j!} \doteq \frac{e^{-1}}{R!}.$$

The first n cumulants of $p_0(R)$ are the same as the first n cumulants of the Poisson distribution with parameter one. [See, for example, J. O. Irwin (1955)].

Spearman's ρ may be written

$$\rho = 1 - \frac{6}{n(n-1)(n+1)} \sum_{i=1}^{n} (r_i - i)^2 = \frac{12S}{n(n-1)(n+1)} - \frac{3(n+1)}{n-1},$$

where $S = \sum_{i=1}^{n} ir_i$. Given that there are R matches, the probability that $r_i = i$ is seen to be

$$\{{}^{n-1}C_{R-1}\ \underline{||n-R}\}/{}^{n}C_R\ \underline{||n-R},$$

which equals R/n, where we use Whitworth's sub-factorial notation,

$$\underline{||m} = m! \sum_{j=0}^{m} \frac{(-1)^j}{j!}.$$

Again the probability that $r_i = k$, where $k \neq i$, is [1]

$$\{{}^{n-2}C_R\ (\underline{||n-R-1} + \underline{||n-R-2})/{}^{n}C_R\ \underline{||n-R}\} = \frac{n-R}{n(n-1)},$$

since

$$\underline{||m+1} = m\ (\underline{||m} + \underline{||m-1}).$$

It follows that for fixed R,

$$\mathscr{E}(S|R) = \sum_{\substack{i=1 \\ R \text{ fixed}}}^{n} ir_i = \sum_{i=1}^{n} i \left(\frac{R}{n} \cdot i + \sum_{k \neq i} \frac{k(n-R)}{n(n-1)} \right) = R\mu'_2 + (n-R)\mu'_{11},$$

where μ'_2 and μ'_{11} are the second order crude moments of the first natural numbers. Substituting

$$\mu'_2 = \frac{(n+1)(2n+1)}{6}, \qquad \mu'_{11} = \frac{(n+1)(3n+2)}{12},$$

we have

$$\mathscr{E}(S|R) = \frac{n(n+1)}{12}(R+3n+2).$$

In similar fashion

$$\begin{aligned}\mathscr{E}(S^2|R) = {} & (\mu'_{22} - 2\mu'_{211} + \mu'_{1111}) + R(\mu'_4 - 3\mu'_{22} + 2n\mu'_{211} - 2(n-1)\mu'_{1111}) \\ & + 2n\mu'_{22} + n(n-2)\mu'_{1111} - \alpha_{n-R}(n-R-1)(\mu'_{22} - 2\mu'_{211} + \mu'_{1111}),\end{aligned}$$

where

$$\alpha_{n-R} = \frac{\sum_{j=0}^{n-R-1} \frac{(-1)^j}{j!}}{\sum_{j=0}^{n-R} \frac{(-1)^j}{j!}},$$

and the μ's are the crude moments of order 4 of the first n natural numbers. These are

$$\mu'_4 = \frac{(n+1)}{30}(6n^3+9n^2+n-1), \qquad \mu'_{22} = \frac{(n+1)}{180}(20n^3+24n^2-5n-6),$$

$$\mu'_{211} = \frac{(n+1)}{360}(30n^3+35n^2-11n-12), \qquad \mu'_{1111} = \frac{(n+1)}{240}(15n^3+15n^2-10n-8).$$

Simple substitution gives

$$\mathscr{E}(\rho|R) = \frac{R-1}{n-1},$$

$$\mathscr{E}(\rho^2|R) = \left(\frac{R-1}{n-1}\right)^2 + \frac{1}{n-1} - \frac{[\,2(2n^2+6n+3) + R(n^2-15n-18) + 2R^2(2n+3)\,]}{5n(n+1)(n-1)^2}$$

$$+ (1-\alpha_{n-R})\frac{(n-R-1)(5n+6)}{5n(n+1)(n-1)}.$$

[1] Compare W. A. Whitworth, *Choice and Chance*, Proposition XXXI and Corollary.

It will be noticed that the regression of ρ on R is linear, an interesting fact which we do not use here. If desired, a conditional test of significance becomes possible for sequences of moderate length by assuming that $\rho|R$ is normally distributed. For long sequences the conditional test will not be of great utility, since

$$\mathscr{E}(\rho|R) \to 0 \text{ as } n \to \infty.$$

Further, as the average value of var $(\rho|R)$ is $(n-2)/(n-1)^2$, little precision is gained by the use of the conditional test.

3. To save dual argument we shall write the probability distribution of R as $p(R)$, distinguishing the null and the alternative hypotheses by $p_0(R)$ and $p_1(R)$ where necessary. The same argument will then serve for both cases. We have

$$p(S) = \sum_R p(S, R) = \sum_R p(R) \,.\, p(S|R),$$

whence

$$\mathscr{E}(S) = \sum_S S \,.\, p(S) = \sum_S \sum_R S \,.\, p(R) \,.\, p(S|R) = \sum_R p(R) \,.\, \mathscr{E}(S|R).$$

If $\mu_t'(R)$ denotes the tth crude moment of the distribution, $p(R)$, then, assuming that different rankings which give rise to the same number of matched pairs are equally likely,

$$\mathscr{E}(S) = \frac{n(n+1)}{12}(\mu_1'(R)+3n+2)$$

and

$$\mathscr{E}(\rho) = \frac{1}{n-1}[\mu_1'(R)-1].$$

When $p(R)$ is the distribution of R under the null hypothesis, the mean is unity, and the mean value of ρ is zero as expected. Again

$$\mathscr{E}(S^2) = \mu_2'(R)[\mu_{22}'-2\mu_{211}'+\mu_{1111}']+\mu_1'(R)[\mu_4'-3\mu_{22}'+2n\mu_{211}'-2(n-1)\mu_{1111}']$$
$$+2n\mu_{22}'+n(n-2)\mu_{1111}'-\sum_R p(R)a_{n-R}(\mu_{22}'-2\mu_{211}'+\mu_{1111}');$$

and similarly for ρ^2. The last sum does not appear reducible in general. It may easily be verified for the null hypothesis that it is equal to $(n-2)(\mu_{22}'-2\mu_{211}'+\mu_{1111}')$.

4. $p_0(R)$ is the distribution of R under the null hypothesis, which asserts that in the bivariate population generating the sample, the correlation between x and y is zero. If in fact there is a correlation between x and y, then there will tend to be more matched pairs than might be expected on the null hypothesis. For instance, in the extreme case all the pairs would be matching for a correlation of 1. We propose therefore for the alternative hypothesis that there has been an increase in the number of matchings (different rankings which give rise to the same number of matched pairs remaining equally likely); and we write

$$p_1(R) = p_0(R)[(n-1)\omega+1]^R E_n^{-1}$$

where

$$E_n = \sum_{j=0}^{n} \frac{[(n-1)\omega]^j}{j!},$$

and ω is a parameter of departure from the null hypothesis. Putting ω equal to zero gives us the distribution $p_0(R)$. The exact moments of $p_1(R)$ are perhaps best given in their factorial form. It may be shown that the mth factorial moment of R under H_1 is

$$\mu_{[m]}' = ((n-1)\omega+1)^m \frac{E_{n-m}}{E_n}.$$

Substituting the appropriate moments in the expressions of the previous paragraph and reducing, we find, using ρ_1 and ρ_0 to indicate ρ under H_1 and H_0 respectively,

$$\mathscr{E}(\rho_1) = \omega+\left(\omega+\frac{1}{n-1}\right)\left(\frac{E_{n-1}-E_n}{E_n}\right),$$

$$\text{var } \rho_1 = \frac{1}{n-1}\left\{1+\frac{4}{5}\cdot\frac{n+2}{n+1}\omega-\frac{2}{5}\frac{(n-1)(2n+3)}{n(n+1)}\omega^2\right\}+\Delta_n$$

$$= \text{var } \rho_0+\frac{4}{5}\frac{n+2}{n^2-1}\omega-\frac{2}{5}\frac{(2n+3)}{n^2+n}\omega^2+\Delta_n,$$

where $\Delta_n = \frac{1}{5}\left(\frac{E_n-E_{n-2}}{E_n}\right)$

$$\times\left\{\frac{\omega^3(n-1)(5n^2+9n+6)}{n^2(n+1)}-\frac{\omega^2(5n^2-7n-9)}{n(n+1)}+\frac{\omega(5n^2-18n-32)}{n(n+1)}-\frac{2(5n+6)}{n(n^2-1)}\right\}$$

$$-\left(\frac{E_n-E_{n-1}}{E_n}\right)^2\left(\omega+\frac{1}{n-1}\right)^2.$$

Both the terms $\left(\frac{E_{n-1}-E_n}{E_n}\right)$ and $\Delta_n$can be shown to be negligible by using *Biometrika Tables* (1954), Table 7. For it is easily seen that $e^{-\omega(n-1)}E_n = 1{\cdot}0000$ for $n \geqslant N(\omega)$, where $N(\omega) = 3$, 4, 8, 13, 21, 34 at $\omega = 0{\cdot}05$, $0{\cdot}1$, $0{\cdot}2$, $0{\cdot}3$, $0{\cdot}4$ and $0{\cdot}5$ respectively.

5. Having established the possibility of such an approximation as that just given, the moments of ρ under the alternative hypothesis follow more easily by noting that

$$p_1(R) \doteqdot \frac{[(n-1)\omega+1]^R e^{-[(n-1)\omega+1]}}{R!},$$

and

$$p_0(R) \doteqdot \frac{e^{-1}}{R!}.$$

The approximation, shown to be negligible for the numerical values quoted, lies in assuming that

$$\sum_{R=0}^{n} p_1(R) \doteqdot \sum_{R=0}^{\infty} p_1(R) = 1.$$

The cumulants of $p_1(R)$ are all $(n-1)\omega+1$. If we substitute and reduce, we have

$$\mathscr{E}(\rho_1) \doteqdot \omega$$

$$\text{var } \rho_1 \doteqdot \text{var } \rho_0+\frac{4}{5}\frac{(n+2)}{n^2-1}\omega-\frac{2}{5}\frac{(2n+3)}{n^2+n}\omega^2 \qquad \text{(A)}$$

as before. A further approximation which has been found to be adequate is to write

$$\text{var } \rho_1 = \frac{1}{n-1}\left(1+\frac{4}{5}\omega-\frac{4}{5}\omega^2\right).$$

6. ρ is approximately normally distributed under H_0; and for small departures from H_0 the same assumption will be adequate. Accordingly, to calculate the power of Spearman's ρ to detect the fact that, in the bivariate population generating the sample, there is a positive correlation ω, we find $\bar{\rho}$ and $\underline{\rho}$ from

$$\int_{-\infty}^{\underline{\rho}} p_0(\rho)d\rho = \frac{\epsilon}{2} = \int_{\bar{\rho}}^{\infty} p_0(\rho)d\rho,$$

where ϵ is at choice. The power, $1-\beta$, is given by

$$\int_{-\infty}^{\underline{\rho}} p_1(\rho)d\rho+\int_{\bar{\rho}}^{\infty} p_1(\rho)d\rho = 1-\beta,$$

where $p_1(\rho)$ is here the normal distribution with mean ω and variance as given in (A). Table I gives a few selected values of the power of the test for a range of values of n and ω over which (A) is allowable.

TABLE I. POWER OF SPEARMAN'S ρ TO DETECT A POSITIVE CORRELATION ω

n \ ω	0·0	0·1	0·2	0·3	0·4	0·5
15	0·05	0·077	0·139	0·234	—	—
25	0·05	0·087	0·182	0·327	0·499	—
35	0·05	0·099	0·230	0·351	0·634	0·808

7. In this paper we have described a possible model for the distortion of the random sequence for use as an alternate hypothesis to randomness. The model has been illustrated on Spearman's ρ; but it can be used as an alternative to the random sequence for any bivariate criterion. It is not the only model which might be set up and it will not of course be appropriate in all circumstances. Another possibility, which we hope to discuss on a future occasion, will be to divide the $n!$ sequences into groups corresponding to the score S and to introduce a system of weights for these groups.

REFERENCES

1. Irwin, J. O. (1955). *Journ. Roy. Statist. Soc.*, Series A, CXVIII, 4, 389.
2. Pearson, E. S. and Hartley, H. O. ed.(1954). *Biometrika Tables for Statisticians*, Cambridge University Press.

Digitized by the Internet Archive
in 2026

https://archive.org/details/british-journal-of-mathematical-statistical-psychology_1956-11_9_part-2

Vol. IX Part I The British Journal of Statistical Psychology November 1956

DIFFERENCES BETWEEN TESTERS IN TERMAN-MERRILL TESTING

By W. CURR and N. GOURLAY
University of Birmingham

I. Purpose of the Investigation. II. Experimental Design. III. Statistical Analysis of the Results. IV. Summary and Conclusions.

I. PURPOSE OF THE INVESTIGATION

In the field of mental testing a common source of error which is often overlooked is the difference, both in establishing rapport and in assessing test performance, between the persons who apply the tests. This shows itself in two ways: first, the test results obtained by a particular tester may have a systematic or mean error, i.e., they may be higher or lower on the average than those obtained by other testers; secondly, testers may differ in the amount of random error which affects their measurements. The present study was undertaken to examine and compare the magnitude of such errors, both mean and random, incurred by trained and untrained persons using the Terman-Merrill tests.

II. EXPERIMENTAL DESIGN

The trained testers consisted of eight educational psychologists, all (with one exception) members of the Committee of Professional Psychologists; the untrained testers, of ten students taking the Birmingham Diploma course in the Psychology of Childhood, all experienced teachers, but with no degree in psychology and no experience of Terman-Merrill testing. Both groups were informed that the purpose of the investigation was to study the amount of error involved in such testing; but neither was aware of the participation of the other group.

The plan of the experiment was to arrange the testers within each group in pairs. The pairs were then allocated to different primary schools. In each school ten boys and ten girls within the age range 8 years 3 months to 9 years 2 months were selected at random. In a first testing, ten children (five boys and five girls) were tested by one tester and ten by the other. At a second testing, approximately three months later, the testers exchanged groups, i.e., tested the ten children previously tested by the other. For both first and second testing the same version—Form L—had to be used, since few of the trained psychologists had had sufficient experience with Form M. Approximately two months after the first testing, all the pupils were tested with Raven's Matrices and Richardson's Simplex Junior. This was carried out by a single person who had no share in the Terman-Merrill testing.

III. STATISTICAL ANALYSIS OF THE RESULTS

We shall use the letters A, B, C, D and E to denote the five pairs of untrained testers and the schools in which they tested, and the letters P, Q, R and S to denote the four pairs of trained testers and the corresponding schools. The experimental design necessitated an analysis of variance of the I.Q.'s for each school. The analysis for school R, shown in Table I, will serve as illustration. The variance *within testees* represents the random error for the two testers. Obviously the statistical test for any mean error difference between the testers is made by comparing the *testers*' variance with the variance *within testees*. Practice effect, if any, enters into the variance between *periods*, which would also be compared with the variance *within testees*. The variance for *sequences* arises from the fact that the two groups

into which the testees were divided were not tested in the same order by the testers: one group was tested first by tester X and then by tester Y, the other by tester Y and then by tester X. This variance can be compared with the variance *between testees*; but the precision of the test is low. In this experiment the differences always proved statistically non-significant. And indeed on *a priori* grounds one might expect that differences between sequences would be negligible.[1]

The nine analyses were also carried out in terms of the logarithms of the I.Q.'s (i.e., 100 $\log_{10}$ I.Q.), where the logarithm was taken to two decimal places. As McNemar has

TABLE I. ANALYSIS OF VARIANCE FOR SCHOOL R (I.Q.'s)

Source of Variance	Degrees of Freedom	Sums of Squares	Mean Square
Testers	1	224·45	224·45
Periods	1	938·45	938·45
Sequences	1	510·05	510·05
Residual { Between testees	16	11,613·20	725·83
Residual { Within testees	16	372·80	23·30
Total	35	13,776·75	—

TABLE II. ANALYSIS OF DIFFERENCES IN MEAN ERROR

Schools	Mean Error Differences (I.Q.'s)	Sums of Squares (I.Q.'s)		Significance (P for F-ratio)	Practice effect (I.Q.'s)
		Between Testers	Random Error (with degrees of freedom)		
A	2·1	38·75	274·78 (15)	·172	−0·6
B	(−)1·2	13·23	485·06 (18)	·493	5·2
C	0·0	0·02	370·42 (15)	·999	4·2
D	1·8	28·26	562·21 (15)	·399	3·0
E	2·1	42·67	250·92 (17)	·110	5·7
P	(−)0·2	0·40	258·50 (18)	·869	3·9
Q	1·1	10·03	337·11 (16)	·500	8·6
R	5·0	224·45	372·80 (16)	·006	9·7
S	0·2	0·23	344·66 (18)	·913	4·8

pointed out ([1], pp. 44-47), the random component of the error of measurement tends to be proportional to the magnitude of the I.Q. Homogeneity of error variance is restored by converting to logarithms before applying the analysis ([2], pp, 198, 199).

1. *Differences between Testers.* The statistics necessary for the study of differences between testers are tabulated in Tables II and III.

(*a*) *Mean Error Differences.* The mean error difference for a pair of testers is the difference between the mean I.Q's which the two testers would obtain for a large sample of testees. An estimate of the difference was derived for each pair, and is given in col. 2 of Table II. As already stated, a

[1] At nearly all the schools one or more of the children were absent at the second testing. As a result, an analysis of variance suitable to unequal (and disproportionate) numbers in the sub-classes was applied for certain schools. This explains why the *total* sum of squares for school R is not equal to the sum of the separate sums of squares.

statistical test of the significance of these differences was carried out in the usual way; and the results are shown in col. 5. Only the difference for school R reaches a high level of significance (0·6 per cent.). The problem now arises of combining these separate analyses.

An application of Pearson's (P, λ_n)-test [3] or its better-known equivalent, the chi-squared test ([4], pp. 99-101) gives a value of P of 0·181, which is far short of significance even at the 5 per cent. level. But since the test utilizes only the P-values of the nine analyses, its precision will be lower than that of any other valid test which makes fuller use of the data. Such a test is available if we can assume approximate homogeneity of random error variance for the nine tester pairs. Homogeneity is by no means perfect; but, as judged by Bartlett's test, the amount of heterogeneity must be small; and the test now to be described is not likely to be seriously biased.

The test is carried out by adding the sums of squares between testers and dividing by nine to give a mean variance between testers, and then repeating the procedure for the sums of squares within testees (dividing in this case by 148), and comparing the mean variance between testers with the mean variance within testees. The F-ratio proves to be 1·86—a figure not far short of significance at the

TABLE III. VARIANCES OF RANDOM ERRORS

Schools	Mean I.Q.	Error Variance (I.Q.'s)	Corrected Error Variance	Error Variance (log I.Q.'s)
A	90·3	18·3	22·5	4·07
B	106·3	27·0	23·8	4·63
C	87·9	24·7	32·0	5·12
D	98·6	37·5	38·6	7·52
E	87·7	14·8	19·4	4·30
P	94·2	14·4	16·4	3·37
Q	104·3	21·1	19·3	3·99
R	100·4	23·3	23·1	4·31
S	111·9	19·2	15·6	2·37

5 per cent. level. When the test was repeated with logarithms of the I.Q's, an F-ratio of 1·96 was obtained, which is just significant at that level; but, as will be seen presently, heterogeneity of random error variance is increased when logarithms are used.

Our data therefore suggest that mean differences exist between testers, though slight, i.e., small systematic errors have occurred in the Terman-Merrill testing. There is, however, nothing in our data to suggest any difference between trained and untrained testers in this respect. An estimate of σ_M^2 (the variance of the tester means) is given by the equation: mean variance between testers = mean variance within testees $+\frac{4}{9}\hat{\sigma}_M^2 \sum \frac{mn}{m+n}$, where m and n represent the number of children in the two groups tested at each school, and the quantity following the summation sign is summed over the nine schools. Thus calculated, $\hat{\sigma}_M^2$ was found to be 1·26 which is quite small in comparison with the mean random error of measurement.

It might be argued that much of the variance, $\hat{\sigma}_M^2$, is due to the difference between the R pair of testers, and that instead of tester means being distributed normally about a ' true ' mean they might have a leptokurtic (and probably skewed) distribution: in that case a knowledge of σ_M^2 would not be very informative. Our data, however, are insufficient to support such an argument.

(*b*) *Random Error Differences.* The random error variances for the nine pairs of testers are shown in col. 3 of Table III. As stated already, Bartlett's test gives no indication of heterogeneity. The error variances of the pairs of trained testers are on the average less than those for the untrained testers; but the difference is obviously not significant. It will, however, be noticed from col. 2 of Table III that the mean I.Q's for the schools in which the trained testers worked are on the whole higher than the means for the five other schools. Hence, if random error is proportional to I.Q. level (as McNemar has shown) some allowance for this must be made in order to make a valid comparison between the two kinds of tester. This can be done by multiplying each variance by the square of the ratio of 100 to the mean I.Q. level, and thus assessing the value of each variance at a mean I.Q. level of 100. The corrected error variances so obtained are shown in col. 4 of Table III. An alternative is to convert the I.Q's to log. I.Q's, and to compare their variances. The latter are given in col. 5. The error variances for the trained and untrained testers now show significant differences. Pitman's test ([5] pp. 128, 129) gives a P-value of 0·048, i.e., the difference is significant at the 5 per cent. level.

While therefore, no difference was found between the trained and untrained testers in regard to mean error, the random error for trained testers would appear to be on an average somewhat less than for untrained testers. At the 100 I.Q. level, the mean random error variance for trained testers works out at approximately 18·2, and that for untrained testers 27·4. Assuming the total variance for 8-9 year-old children to be about 300, these figures correspond to reliability coefficients [1] of 0·94 and 0·91 respectively.

2. *Practice Effect.* Estimates of practice effect for each of the nine schools are shown in col. 6 of Table II. Those recorded for schools Q and R appear excessive. All attempts to find an explanation for this have proved unsuccessful.

3. *Item Comparisons.* A study was made of differences between testers and of practice effect for each Terman-Merrill test item.

(*a*) *Vocabulary Test.* This was treated as a single item. Since it provides a range of scores for the testees, a variance analysis can be carried out. Table IV gives the mean differences between

TABLE IV. ASSESSMENTS OF ERROR AND PRACTICE EFFECTS FOR THE TERMAN-MERRILL VOCABULARY TEST

Schools	Differences in Mean Error		Effects of Practice	
	Words	Months of Mental Age	Words	Months of Mental Age
A	−0·58	− 4·7	−0·08	−0·7
B	−1·45	−11·6	0·85	6·8
C	0·09	0·7	0·53	4·3
D	1·22	9·8	−0·22	−1·8
E	0·23	1·9	0·57	4·5
P	0·25	2·0	1·75	14·0
Q	0·28	2·2	1·06	8·4
R	1·16	9·3	0·54	4·3
S	−0·25	2·0	0·85	6·8

testers and the practice effect for each of the nine schools. These are stated in terms both of words passed and of months of mental age (three words being equivalent to two years of mental age).

The mean differences between testers for schools B, D and R are significant at the 1, 3 and 4 per cent. levels respectively. All three differences are in the same direction as the mean error differences (compare Tables II and IV). Since approximately three months elapsed between the first and second testing, the figures for practice effect given in Table IV should be reduced by about three months of mental age to allow for maturation.

(*b*) *Other Terman-Merrill items.* Since responses to these items are scored + or −, we cannot apply the variance analysis so far employed. Instead, the following procedure was adopted. Since each testee is tested twice, once by each of two testers, his two responses to an item will place him in one of four categories (− −), (− +), (+ −), and (+ +). Thus, if he is in category (− +) he fails the item when tested by the first tester and passes when tested by the second. If, now, m testees are tested first by tester X and then by tester Y and n testees are tested in the reverse order, the $(m+n)$ testees can be distributed according to a 2×4 frequency table, namely:

	(− −)	(− +)	(+ −)	(+ +)
X followed by Y	a	b	c	d
Y followed by X	e	f	g	h

where the sum of the cell frequencies a, b, c and d is m, and the sum of the other four frequencies is n.

[1] These coefficients were obtained by using the formula

$$r = 1 - \frac{\text{error variance (for group)}}{\text{total variance (for group)}}.$$

(Cf. Burt, ' Test Reliability estimated by Analysis of Variance ', this *Journal*, VIII, p. 106, eqn. v). They are not, of course, reliability coefficients for the particular individuals, but rather (to use a phrase suggested by Professor Burt) ' coefficients of intra-group consistency '. This description would accord quite well with Mahmoud's classification (this *Journal*, VIII, pp. 128 f.).

Differences between Testers. A measure of the magnitude of the mean difference between testers is the difference between the two quantities $(b+g)$ and $(c+f)$. To test the significance of such differences, the chi-squared test for 2×2 tables ([4] pp. 96, 97) was applied first to the central 2×2 portion of our 2×4 table and then to the 2×2 table obtained by combining the first, third and fourth columns. The higher of the two levels of significance so obtained was taken as an approximation to the true level. With this procedure differences between $(b+g)$ and $(c+f)$ of 5 and upwards gave levels of significance of 10 per cent. and less.

An examination of the data for differences of this magnitude furnished the following results:

(i) In the case of tester pair R, who, it will be remembered, showed a marked mean difference, no difference was as large as 5. On the other hand, out of 42 test items, as many as 31 showed differences in the same direction as the mean difference between the testers; only 7 items show a difference in the opposite direction. This suggests that whatever factors were operating to produce the significant mean difference between the members of this pair were present throughout the test and not just confined to several items only. Differences in rapport rather than in scoring would appear therefore to be the more likely cause.

(ii) Of the remaining eight pairs, one showed no large difference, five showed one large difference each, and pairs P and S showed three and five respectively, the directions being mixed: thus, although the members of each pair showed significant differences in several of the items, the differences apparently cancelled one another out when an average was taken.

Practice Effect. The size of the practice effect for a given item is indicated by the difference between the two quantities $(b+f)$ and $(c+g)$, after an allowance is made for maturation. The significance of such differences can be tested either by the chi-squared test or by one based on the binomial expansion. There were thirty-three instances of differences of magnitude 5-9, distributed over eighteen test items. The details suggest that certain types of item (e.g., similarities, verbal and picture absurdities and word naming) are more subject to practice effect than others.

4. *Inter-Test Relations.* In order to examine the relations between results for the Terman-Merrill tests and those for the Simplex and Matrices tests, children were selected who had missed none of the tests.' For every school the mean I.Q. was calculated for each of the two Terman-Merrill testers —an allowance being made for practice effect in the case of I.Q.'s obtained at the second testing; and these means were plotted against the corresponding Simplex and Matrices means. Only the diagram for the Simplex test is here reproduced (Figure 1). Strictly speaking, since I.Q.'s involve an age correction, they should not be plotted against scores; but, as the school groups varied but little about the mean chronological age of 8·7 years (at the first testing) no serious error is involved.

The diagrams that we have made (not reproduced here) apparently indicate that the true mean I.Q. of the school populations (i.e., the mean over the total population of testers) is not linearly related to the Simplex and Matrices means. This was confirmed by an analysis of covariance. Hence these diagrams do not (as we had hoped) reveal which of the testers were producing results that were too high or too low.

The use of the Terman-Merrill Results to Standardize Other Tests. Although the relation between the I.Q. means and those for the Simplex and Matrices tests is not exactly linear, it is approximately so. When averages are calculated for the I.Q. means for each tester pair and are compared with those for the Simplex and Matrices, correlations of 0·89 and 0·90 respectively are obtained. Despite the large sampling error, these enable us to check the standardizations of the Simplex and Matrices test at the age level with which we are here concerned. The procedure used consisted in calculating an equation for the equivalence line.[1] For the Simplex test (illustrated in Fig. 1), the equation is

$$(Y-98\cdot2) = 0\cdot917(X-25\cdot1),$$

where Y represents Terman-Merrill I.Q. and X represents Simplex score. At the time of the Simplex testing, the mean chronological age of the children was 8·8 years. Therefore if we take the mean I.Q. for the total population of this country to be 103 (a widely accepted figure), we have for children of the age we are considering, 103 I.Q. units or 106 months of mental age. When Y is expressed

[1] The reason why an equivalence line $(Y-\mu_y) = \frac{\sigma_y}{\sigma_x}(X-\mu_c)$ rather than a regression line should be used in deriving equivalent scores is discussed by Greenall [6]. In a personal communication to one of the authors Professor Burt points out that what Greenall has called the 'equivalence line' is identical with what American writers, following Otis, call the 'relation line' (cf. A. Otis, *J. Educ. Psychol.*, XIII, 1922, pp. 529 f.) and that it is an explicit formulation of the principle commonly adopted by Karl Pearson and his followers, namely, that 'measurements for the same characteristic and the same group of persons, obtained by different methods and expressed in different units, can nevertheless be treated as equivalent, if they are first reduced to terms of their own standard deviation'. Greenall himself observes that similar suggestions, based on the mutual equivalence of scores in standard measure, have been put forward by various writers, in particular Burt (*Distribution of Educational Abilities*, 1917, p. 14) and Thomson (*Psychol. Rev.*, XXXV, 1928, pp. 400-1).

in mental age, the equivalence line becomes $(X-25{\cdot}1) = 1{\cdot}06(Y-101)$. And from this it is easy to show that the Simplex scores equivalent to mental ages of 8, 8·5 and 9 are 19·8, 26·2 and 32·5 respectively.[1] These are much lower than the norms quoted in the *Manual of Directions for the Simplex Test*. Moreover, the discrepancy is too large to be attributed to any inaccuracy in the method. More probably there is a real difference between the population we have tested and the population on which the Simplex test was standardized. There may also be tester differences in the administration of the Simplex test.

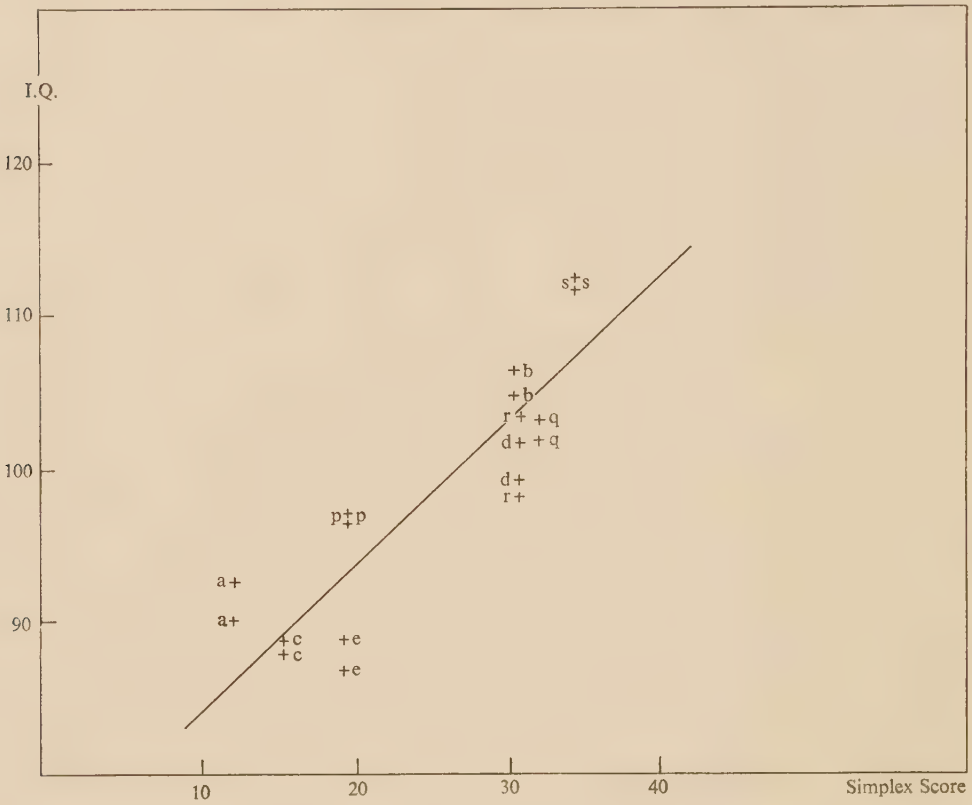


FIG. 1.—Relation between Means of Testers for Terman-Merrill and Simplex Tests.

By a similar method and with the additional assumption that the S.D. of I.Q.'s for the total population is about 17 (see, for example [7]), the following norms were derived for the Matrices test:

		Percentile Points						
		5	10	25	50	75	90	95
Age	8	12	14	17	21	25	28	30
	9	15	17	20	24	28	31	33
	10	18	20	23	27	31	34	36

These figures agree closely with those given in the Matrices *Guide*. Using the same method we obtained the mean Vocabulary test scores for ages of 8, 9 and 10, viz., 7·8, 9·8 and 11·8, respectively. They agree more closely with Terman's and Dunsdon and Roberts' results than with Vernon's ([8] Table V).

[1] This method shows 12·7 units of Simplex score to be equivalent to 12 months of M.A. at this age level, whereas the published norms give a figure of only 8·6 units. The first figure was confirmed when Simplex scores were plotted against C.A. and gave a gradient of 11·9 units of Simplex score per 12 months of age.

IV. SUMMARY AND CONCLUSIONS

1. The purpose of the investigation was to compare the size of the mean and random errors made by trained and untrained Terman-Merrill testers. Eight professional psychologists and ten members of a psychology course took part. Both groups were arranged in pairs. Each pair was assigned to a primary school where they tested a random sample of pupils of chronological age 8-9. There were two testings: at the first, half of each sample was tested by one tester and half by the other; at the second testing, three months later, the testers interchanged groups. The pupils were also tested with the Simplex Junior and Raven's Matrices about two months after the first testing.

2. An analysis of variance was carried out for the I.Q. data obtained by each pair of testers. When the separate analyses were combined, the overall variance between testers was found to be barely significant at the 5 per cent. level. Most of this was contributed by one pair of trained testers, who showed a mean difference of 5 points of I.Q. (significant at the 0·6 per cent. level). It would therefore appear that small systematic errors occur in Terman-Merrill testing. The variance of the mean errors of the testers amounted to 1·26, a small figure in comparison with that for random error. Nothing in the data suggested any difference between trained and untrained testers in regard to mean error.

3. The variance analysis provided estimates of random error for the different pairs of testers. When allowance was made for variation of random error with I.Q., the random error variances of the trained testers proved to be, on the average, less than those for the untrained testers, the difference being significant at the 5 per cent. level. For children of about 100 I.Q., the random error variances of the trained and untrained testers averaged 18·2 and 27·4. These figures are equivalent to reliability coefficients of 0·94 and 0·91 approximately, where the coefficients measure the consistency of two different testers, not the consistency of the same tester on different occasions or with different tests.

4. Practice effect was considerable, and amounted in the case of two of the pairs of trained testers to 8·6 and 9·7 points of I.Q. It was much greater for some items than for others.

5. An analysis was made of the scores for the individual items. In the vocabulary test three tester pairs showed significant mean differences. Several of the other items furnished significant mean differences between testers; but none for the particular pair which showed the large mean difference, although the differences between the members of this pair were nearly always in the same direction as that difference. A difference in rapport rather than in scoring therefore seemed likely.

6. A study was made of the relation between the means of the I.Q.'s and the mean scores in the Simplex and Matrices tests. But the relation was not sufficiently linear to indicate which testers were testing high or low. Equivalence lines were used to derive a set of norms for the Simplex and Matrices tests and for the Terman-Merrill vocabulary test at the age level of the testees.

REFERENCES

1. Terman, L. M. and Merrill, M. A. (1937). *Measuring Intelligence.* London: Harrap.
2. Edwards, A. L. (1950). *Experimental Design in Psychological Research.* New York: Rhinehart.
3. Pearson, K. (1933). ' On a Method of Determining whether a Sample of Size *n* Supposed to have been Drawn from a Parent Population Having a Known Probability Integral has Probably been Drawn at Random.' *Biometrika,* XXV, 379-410.
4. Fisher, R. A. (1946). *Statistical Methods for Research Workers.* London: Oliver and Boyd.
5. Kendall, M. G. (1946). *The Advanced Theory of Statistics II.* London: Griffin.
6. Greenall, P. D. (1949). ' The Concept of Equivalent Scores in Similar Tests.' *Brit. J. Psych. Stat. Sect.,* II, 30-40.
7. Fraser Roberts, J. A. and Mellone, M. A. (1952). ' On the Adjustment of Terman-Merrill I.Q.'s to Secure Comparability at Different Ages.' *Brit. J. Psych. Stat. Sect.,* V. 65-79.
8. Dunsdon, M. I. and Fraser Roberts, J. A. (1953). ' The Relation of the Terman-Merrill Vocabulary Test to Mental Age in a Sample of English Children.' *Brit. J. Stat. Psych.,* VI, 61-70.

Vol. IX
Part II

The British Journal of Statistical Psychology

November
1956

A COMPARISON OF EXACT AND APPROXIMATE LINEAR FUNCTION ESTIMATES OF OBLIQUE FACTOR SCORES

By ANDREW BAGGALEY
Department of Psychology, University of Wisconsin

and RAYMOND B. CATTELL
Department of Psychology, University of Illinois

Abstract.—Exact and approximate procedures for estimating oblique factor scores are described and compared. The amount of error due to the substitution of the approximate procedure is estimated; and the computational problems arising with both methods are briefly discussed.

Problem. Until recently research in factor analysis has for the most part been content to explore factor structure for its own sake. Except in the field of ability, estimates of factor scores for individuals have been rarely made, because evidence for the statistical invariance and psychological meaningfulness of factors was insufficient to justify the use of such scores. However, now that evidence for the stability and psychological meaningfulness of simple structure factors is increasing over a far wider field [2, 5], occasions are likely to multiply, both in research and applied psychology, in which estimates of these factor scores will be required. It will therefore be of interest to workers in educational and psychological measurement to know what are the best computing procedures, to learn something about their requirements in respect of time and instruments, and to consider the merits of approximations.

The present writers have no hesitation in declaring for oblique factors, since their use maximizes the chances of preserving factors with the same psychological meaning from one study to another. Moreover, as Thomson ([8], p. 55) points out, the *estimates* of orthogonal factors are themselves oblique. Both sets of weights—those for orthogonal and those for oblique factors—are used in a simple linear type of equation for purposes of prediction. Hence, once they are determined, the application is equally easy.

An Illustrative Example. A representative example of estimation of oblique factors is available in an area where intensive research has clarified the factor structure. This example involved a confirmatory re-analysis of the Sixteen Personality Factor Questionnaire [4]. 144 new items were packaged in groups consisting of from 3 to 13 items, and each person received a total score on each package, previously examined by item analysis. A sample of 295 undergraduate men and women was scored for 70 package variables. Fifteen factors were extracted and rotated to a clear simple structure, thus supporting the previous interpretations of the variables (excluding general intelligence) in the 16 PFQ.

The exact procedure for estimating oblique factor scores in such a large problem would be impracticable for any laboratory not having access to an electronic computer. However, since the University of Illinois computer ' Illiac ' was available to us, and since such aids are now increasingly accessible to psychologists, it seemed that this study would afford a good opportunity to compare the exact procedure with a simpler, approximate procedure, so that researchers working on smaller problems might have some grounds for deciding which procedure to use.

In a short article Thomson [7] presented a formula for calculating regression coefficients for oblique factor estimates, which involves finding inverses whose order was that of the number of factors instead of the number of variables as in previous formulae ([9], p. 513,

eqn. 11). This is a saving; yet, as Thomson himself admits, the extent of necessary computation is still formidable.

In studies such as that described in this paper, oblique factor estimates are needed to correlate them with further data on the same persons. Since correlations of this type standardize over persons, all that is required are scores which are linear functions of the standard score factor estimates obtained by the complete procedure. Several short cuts thus made possible will be considered below. First, we shall consider exact linear functions and then *approximations* to linear functions.

Notation. We start the development with several definitions:

$N \equiv$ number of persons, (1)

$n \equiv$ number of variables, (2)

$r \equiv$ number of common factors, (3)

$X \equiv N$ by n raw score matrix, (4)

$M \equiv N$ by n matrix of means, all values in a column being identical, (5)

$\sigma \equiv n$ by n diagonal matrix of standard deviations, (6)

$V_o \equiv n$ by r matrix of unrotated common factor loadings, (7)

$V_u \equiv n$ by n diagonal matrix of unique factor loadings, (8)

$\Lambda \equiv r$ by r rotation matrix, (9)

$V_n \equiv n$ by r matrix of rotated reference vector correlations ($V_n = V_o\Lambda$), (10)

$D \equiv r$ by r diagonal matrix whose purpose is to normalize the columns of $(\Lambda')^{-1}$ to make them direction cosines, (11)

$J \equiv [V_o(\Lambda')^{-1}D]'V_u^{-2}[V_nD^{-1}]$, as defined by Thomson [7], (12)

$W \equiv n$ by r matrix of regression coefficients (weights), and (13)

$T \equiv N$ by r matrix of standardized oblique factor estimates. (14)

Note that we have departed from the usual convention ([9], p. 68) in making the score matrices X and T tall and thin instead of short and fat (N should greatly exceed n for stable results). Our transposed arrangement is more convenient for most calculating equipment.

Exact Procedure. Thomson's derivation assumes as an approximation that the factoring reduces the off-diagonal elements of the observed correlation matrix to exactly zero, i.e., that the correlation matrix will equal $V_0V_0' + V_u^2$. With this assumption he derives a formula for the regression coefficients which is, in our notation,

$$W' = (I+J)^{-1}[V_0(\Lambda')^{-1}D]'V_u^{-2}. \tag{15}$$

Here the factor in brackets is the ' primary factor structure ' ([6], p. 278). Then the formula for the standardized oblique factor estimates becomes

$$T = (X-M)\sigma^{-1}W. \tag{16}$$

Multiplying out we have

$$T = X\sigma^{-1}W - M\sigma^{-1}W. \tag{17}$$

Adding $M\sigma^{-1}W$ to both sides gives

$$T + M\sigma^{-1}W = X\sigma^{-1}W. \tag{18}$$

Consideration of the definitions of M and σ will show that all values in a column of $M\sigma^{-1}W$ are identical. Therefore all the entries in a column of T differ from the corresponding entries in the corresponding column of $X\sigma^{-1}W$ by a constant. In other words the columns of $X\sigma^{-1}W$ are exact linear functions of the columns of T. Thus the necessity of standardizing the scores of X is avoided; simply multiply the rows of W by the reciprocals of the standard deviations of the variables (this is analogous to transforming *beta* to *b* coefficients in multiple regression problems), and then premultiply the resulting matrix by X.

If the raw score formula is used for the standard deviation, then it suffices to multiply each row of W by the reciprocal of $N\sigma$ according to the formula, $N\sigma = \sqrt{\{N\sum x^2 - (\sum x)^2\}}$, in which x represents a raw score and the summations are over persons. The resulting

matrix is now $N^{-1}T+M(N\sigma)^{-1}W$ instead of $T+M\sigma^{-1}W$ (N is a scalar matrix [9], p. 21); but the columns of the former are still exact linear functions of the columns of T.

Approximate procedure. So far we have dealt only with *exact* linear function estimates; that is, our estimates for a factor can be transformed into standard score form by merely adding a constant—although the constant will differ from factor to factor. In the remainder of this paper we shall deal with scores which are *approximations* to linear functions of the standard score oblique factor estimates. The empirical results which we present are intended to show how good such an approximation may be.

If the raw scores are dichotomous in form (questionnaire items, for example), and if phi-coefficients are used for the correlation matrix, then $\sqrt{(pq)}$ can be used for the standard deviation, where p is the proportion of high-scored answers and $q \equiv 1-p$. Provided $\cdot 25 < p < \cdot 75$, the reciprocal of $\sqrt{(pq)}$ remains relatively constant, and will not rise more than 16 per cent. above its minimal value. Thus, if dichotomies near the median have been imposed on continuous variables, or if items are used none of which gives a split more extreme than 25: 75 per cent., then W is a close approximation to $\sigma^{-1}W$. Hence the columns of XW will be approximately linear functions of, and highly correlated with, the corresponding columns of T. The same approximation holds in general for continuous variables when their standard deviations differ but little from each other.

However, the computation of W itself presents a more formidable problem. Many research workers have observed in small studies that the reference vector correlations themselves (usually referred to as the factor loadings) are approximately proportional to the regression weights. Accordingly, we decided to investigate how close this approximation would be in a large study.

Using formula (15) a half-time statistical clerk took about six weeks working on the problem with the electronic computer. Of this, however, only a very small proportion was machine time; preparation of tapes and delays in scheduling accounted for the remainder. Programmes for matrix multiplication and the inverse of a symmetric matrix were already available. The need for symmetry entailed slightly modifying the formula. Several of the simpler calculations, such as multiplication by a diagonal matrix, were done on a desk calculator. Since so many multiplications are involved we kept as many as eight decimal places at certain stages.

Inspection of W confirmed that it had the same general pattern as V_n; that is, for a given column the same variables were represented by high entries. However, we desired a quantitative index of the degree of proportionality. We eventually decided on

$$c_j = \sum wv / \sqrt{(\sum w^2 \cdot \sum v^2)},$$

in which w represents an entry in W, v represents an entry in V_n, and the summations are down columns (over variables). This is the product-moment correlation formula when the scores are expressed as deviations from the means. In our problem it represents a kind of regression line which must pass through the origin. It is also identical in form with Tucker's coefficient of congruence ([10], p. 43) and with Wrigley and Neuhaus' index of factorial similarity (unpublished manuscript). We obtained the following values of c_j for the 15 factors and their corresponding reference vectors:

j	c_j	j	c_j	j	c_j
1	·94	6	·90	11	·78
2	·82	7	·88	12	·87
3	·84	8	·82	13	·81
4	·86	9	·94	14	·70
5	·82	10	·89	15	·67

The fit is rather good except for the last two factors. The net result is that a remarkably close approximation to linear functions of the columns of T can be obtained by calculating

$$T_1 = X\sigma^{-1}V_n. \tag{19}$$

First multiply each row of V_n by the reciprocal of the standard deviation of the variable; then premultiply by X. If the standard deviations differ but little from each other (as in the dichotomous situation described above), one can use the formula

$$T_2 = XV_n. \tag{20}$$

Elsewhere the much simpler procedure for estimating scores has been described ([1], p. 80), namely, adding standard scores (or even raw scores) for the highly positively loaded variables and subtracting the same for the highly negatively loaded variables. Such procedures can be represented in the same type of matrix algebra notation used above. However, it should be emphasized that we have no evidence to show how close these approximations are to T.

Form a new matrix $\hat{V}_n$, in which all the high positive loadings of V_n are transformed to values of $+1$, all the high negative loadings to -1, and all the low loadings to 0. Then adding or subtracting *standard* scores gives

$$T_3 = X\sigma^{-1}\hat{V}_n - M\sigma^{-1}\hat{V}_n, \tag{21}$$

which is analogous to (17). However, it is preferable to use the simpler formula,

$$T_4 = X\sigma^{-1}\hat{V}_n, \tag{22}$$

which is analogous to (18). Here, instead of subtracting the mean from the raw score and then dividing by the standard deviation (calculating the standard score), one merely divides the raw score directly by the standard deviation. The columns of T_4 are exact linear functions of the columns of T_3. If the standard deviations differ but little from each other, one can simply add and subtract the *raw* scores according to

$$T_5 = X\hat{V}_n. \tag{23}$$

Formula (23) represents the same approximation to (22) as (20) does to (19).

As a supplement to this discussion, it should be emphasized that *none* of the above short cuts gives factor scores which are comparable from factor to factor. Thus one cannot say whether Mr. X scores higher on Factor A than he does on Factor B. To accomplish this one must have either the standard scores (matrix T) or some linear function of them which is the *same* for all factors. However, the short cut scores are useful for further correlational analysis. The *exact* linear function scores given by formula (18) can also be used for ranking persons on a factor.

Summary. We have described certain exact and approximate linear function estimates of oblique factor scores and discussed the conditions under which they are appropriate. We have shown the extent of approximation involved in using the reference vector correlations (commonly called factor loadings) in place of the exact regression weights in a 15 factor, 70 variable, 295 person problem. We have also discussed some of the computational problems arising when the exact procedure is used.

We wish to acknowledge the aid of Mrs. Alice Joerns who skilfully performed the computations, which were exceedingly complex even with the aid of an electronic computer.

REFERENCES

1. Cattell, R. B. (1952). *Factor Analysis.* New York: Harper.
2. Cattell, R. B. (1955). 'The Principal Invariant Personality Factors Established in Objective Tests.' *J. Abnorm. and Soc. Psychol.*, L, 291-314.
3. Cattell, R. B. 'A Shortened Basic English Version of the 16 Personality Factor Questionnaire.' *J. Soc. Psychol.* (In the press).
4. Cattell, R. B., Saunders, D. R. and Stice, G. (1950). *The Sixteen Personality Factor Questionnaire.* Champaign, Illinois: Institute for Personality and Ability Testing.
5. French, J. W. (1951). *The Description of Aptitude and Achievement Tests in Terms of Rotated Factors.* Chicago: University of Chicago Press.
6. Thomson, G. H. (1948). *The Factorial Analysis of Human Ability* (3rd Edition). Boston: Houghton Mifflin.
7. Thomson, G. H. (1949). "On Estimating Oblique Factors." *Brit. J. Psychol. Stat. Sect.*, II, 1-2.
8. Thomson, G. H. (1954). *The Geometry of Mental Measurement.* London: University of London Press.
9. Thurstone, L. L. (1947). *Multiple-factor Analysis.* Chicago: University of Chicago Press.
10. Tucker, L. R. (1951). *A Method for Synthesis of Factor Analysis Studies.* PRS Report No. 984. Personnel Research Section, Department of the Army.

Vol. IX
Part II

The British Journal of Statistical Psychology

November
1956

THE STATISTICAL ANALYSIS OF THE LEARNING PROCESS

II. STOCHASTIC PROCESSES AND LEARNING BEHAVIOUR

By R. J. AUDLEY and A. R. JONCKHEERE
Department of Psychology, University College, London [1]

Abstract. This paper examines the advantages to be gained from a probabilistic analysis of learning behaviour, and presents a simple conceptual scheme from which are derived a number of stochastic processes suitable for this purpose. An endeavour is made to clarify the problems of estimation and goodness of fit which arise when such processes are fitted to experimental data.

I. PROBABILITY CONCEPTS AND THE DESCRIPTION OF LEARNING

The reasons for employing probability concepts in describing learning behaviour may best be understood if we begin by considering the difficulties that are encountered in summarizing results obtained from experiments involving multiple choice. Here the observations usually consist in noting, for each subject, which of the possible responses actually occurred at each trial. For many purposes limited data of this kind may perhaps suffice: often the grosser predictions made from theories of learning require merely that groups of subjects, under different experimental conditions, shall vary in some overall measure of learning, e.g., in the total number of errors made in a fixed number of trials, or in the number of trials preceding a specified sequence of successes. However, such overall measures are, as a rule, quite arbitrarily chosen; and difficulties arise when groups which differ significantly in one criterion of learning show no such difference when some other criterion is adopted. Again, if the index of learning is taken to be the number of trials needed to reach a certain percentage of correct responses (usually 80 to 100 per cent.), it is not easy to decide precisely how many trials should be allowed.

Attempts have been made to compute the probability of obtaining a similar sequence of correct responses by chance alone (cf., Grant [7]). All such attempts, however, assume that, if no learning has occurred, the probability of a subject making a correct response at each trial will be $1/k$, where k is the number of alternatives available at each trial. But no subject can be presumed to give a sequence of responses which are independent of each other in the manner which this assumption demands. A sequence of independent responses can be obtained by tossing a coin or throwing a die; but it provides a very artificial model for describing learning behaviour. What seems to be wanted is some estimate of the probability of the correct response occurring at a given trial; and the learning criteria commonly adopted provide but crude and indirect estimates of this probability.

Moreover, the use of these overall measures of learning ignores many important aspects of the observed behaviour, in particular, the trial-to-trial changes that such behaviour exhibits. In the past attempts to supply some idea of these changes have been made by fitting learning curves to the experimental data. With situations involving multiple choice this can usually be contrived only by pooling the results obtained from a number of subjects, and so constructing an average curve of learning. Gulliksen [8, 9], however, who cites Merrell [10], points out that such a method can easily lead to incorrect inferences: (see also Sidman [13] and Bakan [2]). In particular, it can be shown that the curve fitted to the average of the individual observations at each trial may imply a mathematical function quite unlike those describing the individual observations themselves even when the latter functions are in form the same for all subjects, and differ only in the individual parameters required.

For many purposes the use of a curve of learning may be misleading. To plot the response variable as a function of the trial number, even for a single individual's performance, may easily

[1] Mr. Audley's work was carried out while he was a member of the Medical Research Council Group for the Experimental Investigation of Behaviour.

obscure the fact that the response made at any particular trial is likely to be a function, not merely of the number of the trial, but also of the sequence of responses rewarded or punished during previous trials.

The methodological problems associated with the construction of learning curves, and the establishment of meaningful criteria for the degree of learning achieved at a particular trial, form but two instances indicating the urgent need for a more detailed analysis of the data obtained from individual subjects. It must, however, be admitted that at first sight the performance of a single subject in a multiple-choice situation provides little information, and can be usefully described only by making explicit assumptions about the dependence of later upon earlier choices. Further, such a scheme would still have to employ probability concepts, since it is unlikely that we shall ever be able to predict with certainty the outcome of each and every trial.

These considerations suggest that the description of learning behaviour is best approached by employing stochastic processes, such as those described by Estes, Bush and Mosteller, and more recently in this *Journal* by Burt and Foley. Such processes are explicitly concerned with the probabilities of the various alternative responses occurring at each trial. And with these procedures there will be a close relation between the empirical data and the experimental interpretation of the theoretical terms employed—' rate ', ' efficiency ', and ' ease ' of learning, and the like—which otherwise remain more or less ambiguous.

Although this paper is principally concerned with the description of a sequence of discrete learning-responses, most of the above considerations also apply to other measures of learning, such as response latency; and in an unpublished paper one of us has suggested a method for relating latency to probability of response. Using this method in conjunction with one of the stochastic processes developed in the present paper, it proved possible to provide a single probabilistic description both for the response probabilities and for the latencies observed in a sequence of learning trials secured from a single subject [1].

II. A CONCEPTUAL SCHEME FOR STOCHASTIC PROCESSES

The stochastic processes to be described were developed by considering the ways in which a succession of drawings can be made from an urn containing coloured balls. This representation of abstract probability schemes will be familiar to readers of statistical textbooks: it resembles that adopted by Burt in the previous issue of this *Journal* [3], and differs from that proposed by Bush and Mosteller [4, 5], who employed a system of operators.

Suppose that an urn contains N balls of two colours, namely, r red, and $N-r$ black, and that a ' trial ' consists of drawing one ball at random, replacing it, and then changing the contents of the urn according to the colour of the ball drawn. Suppose further that the number of balls so added in each case is as shown in Table I.

TABLE I. FUNDAMENTAL URN SCHEME

Ball Drawn	Balls Added		
	Red	Black	Total
Red	a	c_1-a	c_1
Black	b	c_2-b	c_2

We now compute the conditional probability of drawing a red ball at the $(n+1)$th trial, when the results of the previous n drawings are already known. Let X_i represent the outcome of the ith draw so that if the ball drawn on the ith trial is red, $X_i = 1$; if black, $X_i = 0$. Then the probability that the ball drawn on the $(n+1)$th trial will be red, i.e., that $X_{n+1} = 1$, will be

$$\Pr\{X_{n+1} = 1 \mid X_1, X_2, X_3, \ldots X_n\} = \frac{r+(a-b)\sum_{i=1}^{n} X_i + nb}{N+(c_1-c_2)\sum_{i=1}^{n} X_i + nc_2}. \qquad (1)$$

Put $\frac{r}{N} = \rho$, $\frac{a}{N} = \alpha$, $\frac{b}{N} = \beta$, $\frac{c_1}{N}\gamma$, $= \frac{c_2}{N} = \gamma_2$, and $\sum_{i=1}^{n} X_i = K_n$.

On cancelling N, the expression for the required probability becomes

$$\Pr\{X_{n+1} = 1 \mid K_n\} = \frac{\rho+(\alpha-\beta)K_n+n\beta}{1+(\gamma_1-\gamma_2)K_n+n\gamma_2}. \tag{2}$$

Since this is a probability, it must lie between zero and unity inclusive. Hence the parameters must satisfy the following inequalities for all values of n under consideration:

$$1+(\gamma_1-\gamma_2)K_n+n\gamma_2 \geqslant \rho+(\alpha-\beta)K_n+n\beta \geqslant 0. \tag{3}$$

Any failure to fulfil these inequalities would mean that all the balls of one colour had been removed from the urn; and at this point it would be necessary to specify how future drawings were to be made by giving new rules of procedure. For our present purposes, however, we shall assume that these inequalities are not infringed.

Now, let the drawing of a red ball in the urn experiment represent the performance of a correct response in an experiment on learning where only two alternatives are available and where a reward is given each time the response is correct; and let the drawing of a black ball represent the performance of an incorrect and unrewarded response. This elementary situation, in which there is consistent reinforcement, is chosen for simplicity of exposition: we have, however, been able to extend the scheme to cover both partial reinforcement and different training procedures. With the interpretation suggested, eqn. (2) will express the probability of a correct response at the $(n+1)$th trial in terms of the total number of successes and failures which have occurred during the preceding n trials. And, on referring to Table I, it will be seen that the parameters in this expression have equally simple interpretations. In particular ρ will denote the probability of a correct and rewarded response occurring on the first trial, while the values of α, β, γ_1, and γ_2, will indicate the relative influence of reward and punishment upon the subsequent probability of success.

Modifications of the Fundamental Scheme. The basic urn-scheme may be modified in various ways; and we shall give two illustrative examples. In the first we assume that the influence of rewarded and unrewarded responses is restricted to the outcomes of the previous m trials only. This merely affects the summation of X_i, which is now carried out from $i = n-m+1$ to $i = n$. For this procedure we obtain

$$\Pr\{X_{n+1} = 1 \mid X_1, X_2, \ldots X_n\} = \frac{\rho+(\alpha-\beta)\sum_{i=n-m+1}^{n} X_i+m\beta}{1+(\gamma_1-\gamma_2)\sum_{i=n-m+1}^{n} X_i+m\gamma_2}. \tag{4}$$

Inequalities of the kind given by eqn. (3) are again required. But the process may also be regarded as forming a Markov chain, a brief account of which has been given by Burt and Foley in the paper cited above. In such a case the events in the chain will be the 2^m possible sequences of successes and failures in m trials. This treatment has special advantages from a mathematical and statistical standpoint, since the problem of parameter estimation is considerably simplified for small values of m. Another way in which we might modify assumptions as to the effects of previous events on the probability of success at the $(n+1)$th trial, is to weight the effects with some function of their temporal distance from the trial in question. This is the most general case of the urn scheme. In the present paper, however, we shall confine ourselves to the process arising when we assume a geometric progression for the values of the weights. This, as we shall show later, is of special interest in connection with the Bush and Mosteller models for learning. The expression for the probability of success at the $(n+1)$th trial, assuming a weighting of r^{n-i} for the effect of the outcome of the ith trial, is

$$\Pr\{X_{n+1} = 1 \mid X_1, X_2, \ldots X_n\} = \frac{\rho+(\alpha-\beta)\sum_{i=1}^{n} X_i r^{n-i}+\sum_{i=1}^{n} r^{n-i}\beta}{1+(\gamma_1-\gamma_2)\sum_{i=1}^{n} X_i r^{n-i}+\sum_{i=1}^{n} r^{n-i}\gamma_2}. \tag{5}$$

It can be seen that when $r = 1$ we have again the basic unweighted process given by eqn. (2). If r is less than unity, then recent responses have a greater influence on the probability of success than more remote ones. Other modifications may similarly be introduced, but the two just given should illustrate the flexibility of the fundamental urn scheme presented here.

Comparison with other models. Special cases of this scheme will be found to yield models identical with those proposed by previous writers. For example, if in eqn. (2) we put $\alpha = \gamma_1$, $\beta = 0$, then

$$\Pr\{X_{n+1} = 1 \mid K_n\} = \frac{\rho+K_n\alpha}{1+K_n(\alpha-\gamma_2)+n\gamma_2} \tag{6}$$

and we obtain the model described byGulliksen [8] based on an earlier suggestion of Thurstone [14]. It should be pointed out, however, that Gulliksen's development of this scheme obscured its stochastic origin, and that he derived a non-stochastic equation for learning curves. Pólya's urn scheme for describing ' contagious ' processes ([11] e.g., the spread of diseases) is obtained by putting $\gamma_1 = \gamma_2 = \alpha$, $\beta = 0$. It is of greater interest, however, to compare our urn schemes with the stochastic processes recently devised by Bush and Mosteller, who have presented several alternative models in which probability concepts are used to develop a system of operators [4, 5]. The operators state the probability of success in any given trial as a linear function of the probability of success on the previous trial. Thus in a two-choice situation the two operators, Q_1 and Q_2, are given by $Q_i p_n = a_i + \alpha_i p_n$, where i takes the value of 1 or 2 according as the nth trial resulted in success or failure, p_n is the probability of success at the nth trial, and $p_{n+1} = Q_i p_n$. For comparison with Bush and Mosteller any of the models described in this paper may be expressed in terms of operators. Thus for the basic process (2) we find

$$Q_1 p_n = \frac{a+(b+cn)p_n}{(d+cn)+ep_n}, \tag{7}$$

where

$a = (\alpha-\beta)a$, $b = (\alpha-\beta)-\alpha(\gamma_1-\gamma_2)-\rho(\gamma_1-\gamma_2)$, $c = -(\beta\gamma_1-\alpha\gamma_2)$, $d = (\alpha-\beta)+(\alpha\gamma_1-\beta\gamma_2)-\rho(\gamma_1-\gamma_2)$,

and $e = -(\gamma_1-\gamma_2)\gamma_1$. For reasons of space the corresponding Q_2 operator will not be discussed here. It will be noted that the operators we have employed are not linear, like those of Bush and Mosteller, but are functions of the trial number n. A further characteristic of the process is that corresponding operators are commutative, i.e., the order in which they are applied is immaterial. This is a property shared by certain cases of the urn scheme described by eqn. (2), but does not hold for some of the modifications introduced above.

An urn scheme corresponding to the linear operators (7) employed by Bush and Mosteller is given in Table II, where R_n and B_n denote the number of red and black balls in the urn prior to the $n+1$th drawing.

It will be noted that the number of balls of a given colour added at each trial is a function of the number already present. In the present instance, starting from either urn scheme or operator concepts, we can readily produce stochastic processes of a simple nature; but each can only be expressed in terms of the alternative concept with some difficulty.

TABLE II. URN SCHEME WITH LINEAR OPERATORS

Ball Drawn	Balls Added		Total
	Red	Black	
Red	$a_1B_n-(1-a_1-\alpha_1)R_n$	$(1-a_1-\alpha_1)R_n-a_1B_n$	0
Black	$a_2B_n-(1-a_2-\alpha_2)R_n$	$(1-a_2-\alpha_1)R_n-a_2B_n$	0

However, the two approaches lead to the same process if we simplify the urn schemes by putting, in eqn. (2) $\gamma_1 = \gamma_2 = 0$. Then

$$\Pr\{X_{n+1} = 1 \mid K_n\} = \rho+K_n(\alpha-\beta)+n\beta \tag{8}$$

or in terms of operators, $Q_1 p_n = p_n+\alpha$, $Q_2 p_n = p_n+\beta$. The corresponding urn scheme is given by Table I when $c_1 = c_2 = 0$.

The weighted stochastic process (5) leads to a non-commutative model and operators much closer to those employed by Bush and Mosteller, since we obtain their special case when $\alpha_1 = \alpha_2 = \alpha$ in eqn. (7). We have in fact as corresponding operators, $Q_1 p_n = a_1+\alpha p_n$, $Q_2 p_n = a_2+\alpha p_n$. This model is of special interest because Bush and Mosteller derive it from a contiguity theory of learning. However, the urn scheme from which it develops approximates much more closely to a reinforcement theory of learning, with the special assumption that recent rewards and punishments have a greater influence than those occurring earlier.

The fact that different theoretical backgrounds lead to the same stochastic process indicates the difficulty of assigning a given process to any one particular theory, except perhaps on the basis of a somewhat ambiguous criterion of simplicity. The converse problem of finding the stochastic process appropriate to any given theory presents similar difficulties. Additional assumptions are usually needed to formulate such a process, and may lead to very different representations of the same fundamental theory.

The Expected Probability of Success. Many authors have supposed that important features of a stochastic process are revealed by the moments of distribution of the probabilities of success at each trial; and by means of such moments several experimenters [6, 12, 15] have fitted stochastic processes to their data. We shall argue later, when considering the problem of estimating parameters, that these moments are of little value, and that it would be more appropriate to consider the likelihood of the various sequences of responses.

The expectation of the probability of success at the $(n+1)$th trial may be obtained as follows. Let the results of the first n trials be represented by the sequence $(X_1, X_2, X_3, \ldots X_n) \equiv S_n$. Since each trial may result in either success or failure, there are 2^n possible sequences, S_n. Associated with each is the probability of its occurrence, say $P(S_n)$. Further, each also leads to a probability of success at the $(n+1)$th trial, viz., $P(X_{n+1} = 1 \mid S_n)$. The expected probability, $\bar{P}_{n+1}$, is therefore given by $\bar{P}_{n+1} = \sum P(S_n)P(X_{n+1} = 1 \mid S_n)$, where the summation is made over all 2^n possible values of S_n. In order to derive an expression for $\bar{P}_{n+1}$ for the basic process (2), we have been forced to simplify the scheme a little by putting $\gamma_1 = \gamma_2 = \gamma$. For this case we have [1]

$$\bar{P}_{n+1} = \left\{\rho + \frac{\beta}{1+\alpha-\beta}\left[1 + \frac{1+\gamma}{(1+\alpha-\beta+\gamma)} + \frac{(1+\gamma)(1+2\gamma)}{(1+\alpha-\beta+\gamma)(1+\alpha-\beta+2\gamma)} + \ldots \right.\right.$$
$$\left.\left. + \frac{(1+\gamma)(1+2\gamma)\ldots(1+(n-1)\gamma)}{(1+\alpha-\beta+\gamma)(1+\alpha-\beta+2\gamma)\ldots(1+\alpha-\beta+(n-1)\gamma)}\right]\right\} \prod_{K_n=1}^{n} \frac{1+\alpha-\beta+(K_n-1)\gamma}{1+K_n\gamma}. \quad (9)$$

III. THE ESTIMATION OF PARAMETERS

We now consider the application of these stochastic processes to empirical data. Here the greatest difficulties arise; and Bush and Mosteller have only succeeded in estimating the parameters for the simplest processes by introducing a number of simplifying assumptions. Prior to the problem of estimating parameters we must decide which type of process is appropriate to the experimental situation. This turns on the goodness of fit and will be discussed in a moment. Here, however, it must be noted that, if we estimate a single set of parameters from the results of a group of subjects, we are forced to assume that the sequence of responses obtained from each subject exemplifies the same type of stochastic process. On the other hand, if we estimate the parameters for each subject separately, we need no longer assume that the process is the same for all.

Even when we assume one particular type of process to be universally operative, a further source of variation arises from the different parameters characterizing the performances of different subjects. This variation is of the greatest importance in assessing alternative methods of estimation. The estimation of a single set of parameters from group data will usually be more stable, since a greater number of observations are employed. On the other hand, if we estimate the parameters for each subject purely on the basis of his results, we require no further assumptions; whereas, if we attempt to estimate the parameters for the group as a whole, the only practical assumption would seem to be that all display the same stochastic processes with the same parameter values. Such an assumption is so improbable that in general it seems better to estimate the parameters separately for each individual. Furthermore, if we use only one set of parameters to represent the performance of the group, then, although we can study their changes with changed experimental conditions, it nevertheless becomes difficult to determine the statistical significance of such changes.[2] If we use individual estimation procedures, assigning to each subject his own particular set of parameters, then we obtain a sample of values for each experimental condition, which can be compared by standard statistical procedures.

Whichever procedure we decide to follow, we still have to consider whether or not to take into account the interdependence of the observed sequence of responses. To ignore such interdependence is to obscure the very phenomena which we are seeking to describe, and is tantamount to a reversion to the traditional study of ' learning curves '. The foregoing considerations yield four main procedures for estimating parameters, which we shall now describe.

[1] We have obtained formulae for the standard error of P_{n-1}, for the expected probability of success, and for the standard error of the various modifications of the basic process described in eqn. (2). They are omitted here for lack of space.

[2] Standard errors of parameter estimates may be obtained in some cases; but their properties with reference to non-stationary time-series of the type we are investigating are not sufficiently well known to warrant their practical use.

A. Ignoring (*i*) *Interdependence of Observed Responses and* (*ii*)*Individual Differences in Parameters.* The traditional method of fitting curves to learning data falls into this first category. The data obtained from a group of subjects, whether expressed in terms of the average number of correct responses at each trial or in terms of the results from a block of trials, are fitted by a curve which is either selected from a family of curves for which some prior justification is given or based on an appropriate function suggested by the data themselves. In the latter case we have what is commonly called ' blind curve fitting '. In the former case, the justification of the curve may be given in terms either of some theory of learning or of a supposed mechanism of interdependence between successive responses. And with this last standpoint the curve is based on the expected probability of success at each trial (cf. eqn. 9). Estes [6] has employed this technique in his study of partial reinforcement.

B. Ignoring (*i*) *Interdependence of Observed Responses, but Allowing for* (*ii*) *Individual Differences in Parameters.* The work of Gulliksen [8] appears to fall under this heading. Recognizing that individual differences may be an important source of experimental variation, he fits a theoretically derived function of the cumulative numbers of successes and failures in a sequence of responses to the observations obtained from each individual separately. The use of such a cumulative curve, however, does not seem to take account of the interdependence of the observed responses. This is evident from his later paper [9] where an equivalent expression is derived, which gives the probability of a correct response simply as a function of the ordinal number of the trial. Such a procedure clearly ignores the unique response sequence of each individual. Even should the observed cumulative successes of each trial be fitted to the corresponding cumulative function of the average probabilities of success derived from some stochastic process, the particular sequence of responses occurring prior to each trial would be still left out of account.

C. Allowing for (*i*) *Interdependence of Responses but Ignoring* (*ii*) *Individual Differences in Parameters.* This approach is found in Bush and Mosteller [4, 5], where the same parameters are assumed for all subjects. Conceptually, therefore, the individual sequences are regarded as samples taken from the population of all the sequences that could be generated by a stochastic process having a particular set of parameter values. Hence the response of any one subject at any one trial is considered in relation to the specific sequence of responses exhibited by that individual during preceding trials; and estimation is carried out either by fitting theoretical moments or by employing the technique of maximum likelihood. Allowance is thus made for individual differences in experience. Provided we may assume that there is little variation between the true parameters that characterize each subject's individual performances, such a method has obvious advantages, especially in respect of the number of observations on which parameter estimates are to be based. We are, however, bound to ask whether such an assumption is in fact justified; and without some test of homogeneity the validity of the method remains open to question.

The problem of individual differences may be split into two parts. First, there are the differences in the probability of making the correct response on the first trial. In those special cases where the probability is approximately zero for every subject, or, owing to previous training, is approximately unity, no serious difficulty arises. Secondly, there are the differences in the parameters associated with the change in the probability of success as a result of previous successes or failures. Such differences would entail a lack of homogeneity in the response sequences obtained from a group; and this might gravely impair the accuracy of any estimation which ignored such differences. To reveal variations between groups working under different conditions, the method is clearly adequate, although there will still be difficulties in determining whether or not the variations are statistically significant.

D. Allowing for (*i*) *Interdependence of Responses and* (*ii*) *Individual Differences in Parameters.* We ourselves would prefer to advocate the use of methods falling within this last category. They consist in assuming some specific type of stochastic process, and then finding those values for the several parameters which maximize the probability of the particular sequence of responses observed. Accordingly, let X_i represent the response of a subject at the ith trial, X_i taking values of 1 or 0 according as the response was correct or not. Now the probability, p_i, of making a correct response at the ith trial is a function of the previous $(i-1)$ responses, $X_1, X_2, \ldots X_{i-2}$, and the set of parameter values ρ, α, β, γ_1, γ_2, belonging to the particular stochastic process assumed, namely M (say). We then obtain

$$p_i = \Pr\{X_i = 1 \mid X_1, X_2, \ldots, X_{i-1};\ \rho, \alpha, \beta, \gamma_1, \gamma_2;\ M\}. \tag{10}$$

The likelihood, L_n, of obtaining a particular sequence of n responses, $S_n = X_1', X_2', \ldots, X_n'$ will be

$$L_n = \Pr\{S_n \mid \rho, \alpha, \beta, \gamma_1, \gamma_2;\ M\} = \prod_{i=1}^{n} p_i^{X_i'}(1-p_i)^{1-X_i'}. \tag{11}$$

The maximum likelihood estimates $\hat{\rho}, \hat{\alpha}, \hat{\beta}, \hat{\gamma}_1, \hat{\gamma}_2$, are the values of the parameters which maximize L_n. In attempting to obtain these values a number of difficulties arise. To begin with, the problem reduces to finding the solution for a set of simultaneous non-linear equations in several unknowns.

In general these can be found only by some iterative procedure. But that is merely a matter of computation. A more fundamental difficulty lies in the possibility that there may be more than one set of values which will maximize L_n—a difficulty which may in theory arise when estimates are based on data not only for a single individual, but also for a group. For some of the simpler processes, however, it may be demonstrated that there is in fact only one unique set; and the experience we ourselves have so far obtained from a number of computations does not suggest the presence of multiple sets of parameters even for the more general processes.

A simpler though approximate technique, suitable for certain types of stochastic process, may be suggested which will often avoid both the difficulties mentioned, namely, the method of least squares. This is the technique suggested by Burt and Foley, though they express their aim as the computation of a ' principal component ' which is (as usual) defined in the least squares sense. With the least squares procedure we attempt to make the probability of success at the ith trial, p_i, as near to 1 or 0 as possible, according as the result was in fact a success or a failure, and this attempt must be undertaken simultaneously for all the n trials. Thus, if we put $D_n = \sum_{i=1}^{n} (X_i - p_i)^2$, we have to find the set of parameters which minimizes D_n. In practice, although occasionally it yields probabilities larger than unity or less than zero, the method nevertheless has been found to yield useful results.

IV. GOODNESS OF FIT

Having estimated the parameters that characterize a specified stochastic process, we have next to determine whether the properties of this process are in reasonable agreement with the observed course of learning. For a single subject we may compute the likelihood that, given the parameters so estimated, the observed sequence of responses should actually occur.

In assessing the likelihood, we require the distribution of the likelihoods of all the possible 2^n sequences that might have appeared during the first n trials. We should then be able to compare the likelihood of an observed sequence with this theoretical distribution, and ascertain whether the sequence observed was in fact very probable. Such a procedure involves considerable mathematical difficulties; and so far we have been able to obtain only the first three moments of the distribution for the simple case in which $\gamma_1 = \gamma_2 = 0$, and $\alpha = \beta$. Other authors, e.g., Bush and Mosteller [4, 5], have attempted to assess the goodness of fit for group data by employing chi-squared techniques. The method will certainly give some rough idea of the degree of fit; but in general the assumptions underlying the chi-squared procedure will not be satisfied. Hence the probability levels associated with this procedure must be accepted with caution, as Bush and Mosteller recognize.

When no antecedent information is available which will directly indicate the most suitable model, it will, as a rule, be necessary to fall back on some methodological plan such as may enable us to determine which particular process will best describe the sequence of responses actually obtained. Here, therefore, we are reduced to much the same position as the early investigator who was content simply to fit an appropriate curve. However, once a learning model has been adequately fitted, its parameters and their interpretation will be of direct service for the development of a theory of learning, whereas the parameters of a fitted polynomial can seldom claim any clear theoretical meaning. A rational procedure is possible whenever we are able to postulate some general process, i.e., a process from which simpler types may be derived by eliminating some of the parameters or giving them special values. Eqn. (2) offers a concrete example, since it contains as special cases the processes described by eqn. (6) and (8).

Other situations may arise where it is possible on theoretical grounds to postulate some general property of the learning phenomena such as will limit selection to a few specific processes. For example, it might be assumed that rewarding correct responses alone has an influence on learning. With this assumption we might then inquire whether the increment in the probability of a correct response is constant or diminishes as learning proceeds. In the latter case we should obtain the process given by $P(X_{n+1} \mid K_n) = \frac{\rho + K_n\alpha}{1 + K_n\alpha}$; and in the former, $P(X_{n+1} \mid K_n) = \rho + K_n\alpha$. Both arise as special cases of the general model (2): in the first case $\alpha = \gamma_1$, $\gamma_2 = \beta = 0$, and in the second $\gamma_1 = \gamma_2 = \beta = 0$. Which of the two is the more appropriate might be inferred by estimating simultaneously all the parameters of the general process, provided of course that the data are sufficient to permit such an estimation.

The tentative nature of all these suggestions reflects the need for a better theoretical understanding of the statistical properties of non-stationary time series. At present statisticians are mainly concerned with stationary time-series, i.e., the transition probabilities between the various events are assumed to remain constant over the entire sequence. With this restriction the problems of estimation and

goodness of fit will obviously become much simpler. Unfortunately, however, sequences so characterized are by their very nature unsuitable for the exact description of learning behaviour. At present, therefore, it would appear that the best that can be done is to check the adequacy of any technique that may be used by extensive empirical sampling from known populations.

V. SUMMARY

Stochastic processes may be serviceable for the study of learning in two ways: first, to increase precision in formulating and verifying current theories of learning; secondly, to provide a more detailed description of the learning behaviour itself. When description is the primary object, some conceptual scheme is required from which a suitable stochastic process may be deduced. Two such schemes have been used by previous investigators—operator schemes and urn-and-ball (or bag-and-ball) schemes. The present paper attempts a more general examination of schemes of the latter type.

With either type, it is necessary to make adequate estimates of the parameters characterizing the process adopted, and to inquire how far the results derived from these estimates are in reasonable conformity with experimental data. We have discussed in some detail the merits of the various methods available for estimating parameters, and have put forward suggestions for testing goodness of fit.

REFERENCES

1. Audley, R. J. (1957). 'A stochastic description of the learning behaviour of an individual subject.' *Quart. J. Exp. Psychol.* (to appear).
2. Bakan, D. (1952). 'A generalization of Sidman's results on group and individual functions, and a criterion.' *Psychol. Bull.*, XLIX, 263-269.
3. Burt, C. and Foley (1956). 'The statistical analysis of the learning process.' *Brit. J. Statist. Psychol.*, IX, 49-62.
4. Bush, R. R. and Mosteller, F. (1953). *Annals of math. Stat.*, XXIV, 559-585.
5. Bush, R. R. and Mosteller, F. (1955). *Stochastic models for learning*. New York: John Wiley.
6. Estes, W. K. and Straughan, J. H. (1954). 'Analysis of a verbal conditioning situation in terms of a statistical learning theory.' *J. Exp. Psychol.*, XLVII, 225-234.
7. Grant, D. A. (1946). 'New statistical criteria for learning and problem solution involving repeated trials.' *Psychol. Bull.*, XLIV, 272-282.
8. Gulliksen, H. (1934). 'A rational equation of the learning curve based on Thorndike's Law of effect.' *J. Gen. Psychol.*, XI, 395-434.
9. Gulliksen, H. (1955). 'A generalization of Thurstone's learning function.' *Psychometrika*, XVIII, 297-307.
10. Merrell, Margaret (1931). 'The relationship of individual growth to average growth.' *Human Biol.*, III, 37-70.
11. Pólya, G. (1930). 'Sur quelques points de la théorie des probabilités.' *Annales de l'Institut Henri Poincaré*, I, 117-161.
12. Restle, F. (1955). 'A theory of discrimination learning.' *Psychol. Rev.*, LXII, 11-19.
13. Sidman, H. (1952). 'Note on functional relations obtained from group data.' *Psychol. Bull.*, XLIX, 263-269.
14. Thurstone, L. L. (1930). 'The learning function.' *J. gen. Psychol.*, III, 469-493.
15. Wiley, L. E. and Wiley, A. M. (1937). 'Studies in the learning function.' *Psychometrika*, II, 1-20.

Vol. VIII Part II The British Journal of Statistical Psychology November 1956

THE MULTIFACTORIAL THEORY OF INHERITANCE AND ITS APPLICATION TO INTELLIGENCE[1]

By CYRIL BURT and MARGARET HOWARD
University College, London

Οὐκ ἔστ' ἀκριβὲς οὐδὲν εἰς εὐανδρίαν·
ἔχουσι γὰρ ταραγμὸν αἱ φύσεις βροτῶν.
ἤδη γὰρ εἶδον παῖδα γενναίου πατρὸς
τὸ μηδὲν ὄντα, χρηστὰ δ' ἐκ κακῶν τέκνα,
γνώμην δὲ μεγάλην ἐν πένητι σώματι.
πῶς οὖν τις αὐτὰ διαλαβὼν ὀρθῶς κρινεῖ;

EURIPIDES, *Electra*, 367-372.

Abstract. The hypothesis put forward for discussion is that, among normal persons, differences in ' intelligence ' (i.e., innate general cognitive ability) are determined by a large number of genes, segregating in accordance with Mendelian principles and each producing effects that are small, similar, and cumulative. From these assumptions, with a due allowance for dominance and assortative mating, formulae are derived for the correlations to be expected between siblings, parents and offspring, and remoter relatives. Data from school surveys yield coefficients which are in almost complete agreement with the theoretical figures thus deduced. An attempt is also made to analyse the variance of the assessments obtained, and to estimate the contributions of genetic factors and non-genetic factors respectively. Appreciable differences are found between the results obtained from raw measurements based directly on intelligence tests and those obtained from revised or adjusted assessments based on all the available evidence. With the former nearly 25 per cent. of the variance is apparently due to non-genetic factors; with the latter barely half that amount.

I. THE HYPOTHESIS OF MULTIFACTORIAL INHERITANCE

Problem and Definitions. The purpose of this paper is to show that to a close approximation the distributions and correlations obtained for human intelligence, like those obtained for so many other graded characteristics both in animals and in man, conform with the results deducible from the principles of *multifactorial inheritance*, and then, on the basis of this theory, to *analyse the variance* of the assessments available in actual practice and estimate the proportions due to the main contributory factors.

The word ' intelligence ' is used, in accordance with the definition given in previous articles [5, 7], to denote that part of the general cognitive factor which is attributable to the individual's genetic constitution. By ' cognitive ' we understand the directive, adaptive, or cybernetic aspects of conscious behaviour as distinct from the dynamic (or conative) and the valuative (or affective) aspects. If anyone thinks this an unsuitable interpretation of the term ' intelligence ', then let him choose or coin a better word of his own, or substitute some purely arbitrary symbol like γ. What he may alter is only the name, not the concept. To retain the label intelligence, assign it a wholly different meaning, and then show (as so

[1] We are much indebted to Dr. Fraser Roberts and Dr. J. O. Irwin for their kindness in reading an early draft of this paper, and for their comments and criticisms, and to Dr. J. S. Wilkie for his help in reading the proofs. We have also to renew our thanks to the many teachers and research-students who assisted in the collection of the data discussed in the later sections.

many critics have done) that our arguments and conclusions break down when applied to this modified concept, is simply to confuse the issue.

During the recent controversies about the use of intelligence tests in schools many psychologists have questioned the very idea of an innate or inherited factor underlying apparent differences in intelligence. But, with few exceptions, the critics seem to entertain entirely out-of-date notions about the nature of human heredity—notions which they also attribute to the writers whom they criticize. Heredity is assumed to be a matter of *resemblances*, and the genetic determination of *differences* is almost wholly ignored; nor is any clear distinction drawn between the genetic determinants and their ultimate manifestations—between 'genotype' and 'phenotype'. As a result the question is discussed almost exclusively in terms of correlations, instead of in terms of an analysis of variance or an analysis of factors. This somewhat antiquated attitude is virtually that which was adopted over fifty years ago by Karl Pearson and by what was commonly called the biometric school. As Calvin Hall observes, 'psychology, though it has taken a lively interest in heredity, is too often content with methods of pre-Mendelian vintage'.[1] The few who refer to the Mendelian hypothesis still apparently regard it, in company with Karl Pearson and his co-workers, as a rival theory incompatible with the results of correlational techniques; and anyone who advocates a Mendelian interpretation is thought to be committed to the view that intelligence is a 'unit character', or that it is composed of a number of 'primary abilities' which are themselves 'unit-characters', so that the graded variations discernible in everyday life have to be explained as the result of environmental differences.

Biometrical Genetics. So long as the Mendelian theory was limited to a doctrine of unifactor inheritance, it was natural to suppose that its applications could best be studied by experimental methods. But the demonstration of multifactor inheritance implies the existence of genes whose effects cannot be traced individually by Mendelian techniques, but only examined statistically in the light of Mendelian principles. In a memoir that has since become classical, Fisher [17] virtually ended the long conflict between the Mendelians and the biometricians [2] by showing that, on this

[1] Calvin S. Hall, 'The Genetics of Behaviour', *ap.* S. S. Stevens, *Handbook of Experimental Psychology*, 1951, p. 304. Over forty years ago Burt ([1], 1912, pp. 20 f.) tried to emphasize the importance of the Mendelian theory for the problems of psychology. But, as the above quotation from the *Electra* shows, the Greek of the 5th century B.C. was in many ways ahead of the modern psychologist in his reflections on human heredity, and had already recognized that the key was to be found in the study of differences rather than of resemblances. According to Aristotle (*De Generatione Animalium*, 769a 25) some writers sought to account for them by a multifactorial version of the particulate theory (*πανσπερμία τίνα πολλῶν*); but he preferred his own notion of the male parent impressing a 'form' on the 'matter' supplied by the female. The Galenic doctrine of 'humours' led to the alternative view, which held the field till the late 19th century, and indeed still underlies the tacit assumptions of most present-day psychologists—namely, that innate characteristics result from a 'blending' of the 'blood' received from each parent. Even Galton explained his 'regression coefficient' in terms of this metaphor (*Natural Inheritance*, 1889, pp. 105 f.).

[2] The first to suggest using statistical procedures to formulate and test a multifactorial theory was Udny Yule (1902). Pearson (1904) accordingly devoted the twelfth of his *Mathematical Contributions* to an examination of the proposal. He proved, or believed he had proved, that the theory broke down for two reasons: (i) it led to values for the correlations between relatives which, in the light of a wide range of data, were far too small; (ii) it implied that each correlation would have a fixed value no matter how different the species or how different the trait; this also was contradicted by the facts. The earliest attempt to combine Mendelian with biometric principles is to be found in a series of papers by Weinberg [60, 61]; but these have remained almost unnoticed both in Britain and in America. Fisher dealt directly with the difficulties raised by Pearson and his co-workers; and, as we shall see, his treatment furnished more comprehensive formulae and issued in higher values for the correlations, thus overcoming both Pearson's objections. At that time, however, Fisher's methods of handling statistical problems were too new and too technical to attract much attention; and it was for long widely supposed that Pearson's results had conclusively shown the 'utter inapplicability of any Mendelian hypothesis to continuous variations in man'. Thus, for the next fifteen years it was generally agreed by both Mendelians and biometricians that, whichever hypothesis came nearer the truth, the hereditary determination of graded traits was quite incompatible with the segregation of discontinuous units postulated by the Mendelian theory. This view still seems to survive in the arguments adduced by many psychological writers. Nevertheless, as the recent advances of genetics have amply shown, on the one point over which the two disputants happened to agree, both, strange to say, were utterly wrong.

basis, all the known facts regarding the distribution and correlations of graded characteristics (such as height, span, arm-length, etc.) could be readily accounted for in terms of a generalized version of Mendel's original hypothesis. The result was a union of Mendelism and biometrics, and the consequent extension of Mendelian principles to embrace almost all forms of genetic variation.

In several earlier publications, one of us [3, 7 and refs.] has ventured to suggest that a similar mathematical treatment could be applied to mental characteristics, such as intelligence. However, few psychologists seem familiar with Fisher's statistical procedure; and, since the inheritability of differences in intelligence has of late been so frequently challenged, it has been suggested that a simplified exposition of the whole theory, starting from elementary principles and avoiding so far as possible abstruse technicalities, might be of interest to the general reader. Although recent work in biometrical genetics has introduced several modifications into the arguments as originally stated, Fisher's analysis makes an excellent starting-point for discussion and further research, and his basic formulae seem to have gained general acceptance (cf. [11], pp. 80 f.).

The Hypothesis. Let us begin by formulating the hypothesis in general terms. The central assumption is that, so far as genetic influences are concerned, an individual's measurement for intelligence is determined by a large number of independent factors, each of whose effects is *small*, *similar*, and *cumulative*.[1] The mechanism of hereditary transmission we shall assume can be summed up in the following postulates, all of them suggested by phenomena that have been repeatedly verified for a wide range of qualitative characteristics in animals and plants. (i) Each factor, though itself invisible and indeed hypothetical, may be thought of as possessing a *definite location on one of the chromosomes*, the visible but microscopic bodies within the cell-nucleus which form the essential vehicles of inheritance. (ii) Each factor has *two alternative forms* (' allelomorphs '), of which one or the other will be carried at corresponding loci on corresponding chromosomes. For convenience these alternative forms will be designated by large and small versions of the same letter, *A* and *a*, *B* and *b*, and so on. (iii) Every factor *preserves its nature and form* unaltered from one generation to another: it never changes, vanishes, or blends, however often its observable effects may seem to do so. (iv) Every *gamete* (i.e., every germ-cell, whether egg or sperm) contains one and only *one specimen* of each factor, never both: viz., either *A* or *a*, and either *B* or *b*, and so on; every *zygote* (i.e., every fertilized cell formed by the union of two gametes) contains *two specimens* of each factor, one derived from the male parent and one from the female: they may be of different form (*Aa* or *aA*), or both of the same form (*AA* or *aa*). (v) As a result of the separation of the chromosomes during the nuclear divisions which precede the formation of the germ-cells, the two factors thus brought together in the zygote will again be *segregated* in the gametes, the distribution of the factors from each pair being determined purely by chance; and when yet another zygote is formed by fertilization they will be *recombined* in pairs, the type of re-combination being once again determined purely by chance. (vi) The segregation and recombination of each pair occur *in complete independence* of the segregation and recombination of other pairs, so that the final re-assortment is also determined by chance.

The foregoing postulates are not intended to deny the occasional occurrence of cytoplasmic (extra-nuclear) heredity, multiple allelomorphs, mutation, linkage and sex-linkage, and various irregularities in the behaviour of chromosomes: they merely assume that, for purposes of the deductions and computations that are to follow, these complicating possibilities may be temporarily disregarded as negligible or irrelevant. Towards the close of our discussion, however, we shall glance briefly at the effects of removing some of these provisional restrictions. Since in statistical psychology the word ' factor ' is commonly used in a broader and a rather different sense, it will often save ambiguity and circumlocution if we refer to genetic factors briefly as ' genes '.[2]

[1] In speaking of the effects of the different factors as ' similar ', we are referring solely to their effects on the particular character under discussion. Pleiotropy (the initiation by a single gene of several physiologically independent chains of reaction) seems to be a widespread rule in most organisms; recent work by MacArthur and others has furnished impressive lists of such multiple effects: e.g., selecting mice for large and small body-size is apt to bring with it changes in shape, coat-colour, fertility, and even temperament.

[2] The word ' gene ' is an anglicized translation of the German ' Gen ' (plural ' Gene '), coined by Johannsen to designate ' the hereditary element or factor responsible for the phenomena of segregation '; the term, he explains, was suggested by Darwin's word ' Pangenesis '. For Johannsen, however, it denoted, not a concrete entity (whether particle, gemmule, or the like) but a purely

Corollaries. The essential feature of the Mendelian hypothesis lies in the fact that both the segregation of genes and their recombination are, in Mendel's own phrase,[1] *ganz dem Zufalle überlassen* (' wholly determined by chance '). It follows that the average or expected results may, as he pointed out, be predicted *nach den Regeln der Wahrscheinlichkeit* (' according to the rules of probability '). Thus the problems of heredity, even in its most elementary manifestations, are primarily statistical problems.

(*a*) *Unifactorial Inheritance.* In the simpler and more familiar instances of Mendelian inheritance two different forms of the same factor or gene are assumed to be responsible for alternative qualities in the same organ or structure: e.g., if *A* is responsible for producing tall peas, *a* will produce dwarfs; if *B* produces coloured blossoms, *b* will produce white. And with the mechanism described above, the occurrence of any particular gene, *A-a* say, will imply the possibility of four zygotic combinations, viz. (putting the paternal gene first), *AA*, *Aa*, *aA*, and *aa*, and therefore of three distinct types (since except in their origin *Aa* and *aA* are identical). Individuals possessing two similar forms of the same gene (*AA* or *aa*) are called ' homozygous ' (i.e., ' pure '), and those possessing two dissimilar forms (*Aa* or *aA*) are called ' heterozygous ' (i.e., ' hybrid '). It also follows that, if we cross two different homozygotes, say *AA* and *aa*, all the offspring will belong to the hybrid or heterozygous type *Aa*: and, if in turn these offspring are self-fertilized or interbred, their offspring will in the long run include specimens of all four combinations, *AA*, *Aa*, *aA*, and *aa*, in approximately equal numbers. Moreover, in the case of dihybrids, since we have assumed that the reassortment of the different factors in the new zygote is determined by chance, any given combination for one factor, *Aa* say, will in the long run be combined with *BB*, *Bb*, *bB*, and *bb* with equal frequencies. These lead to the wellknown 'Mendelian ratios' which provide such convenient checks.

In adopting a hypothesis of this kind we are evidently bound to recognize, and allow for, a clear distinction between the hypothetical ' genotype ' and the observable ' phenotype ' [28], that is to say, between the genetic characteristics of the individual, resulting from the properties of the genetic material transmitted by his parents, and the characteristics ultimately observed or measured, since the latter will be affected by the interaction of his various genes with each other, with the rest of the soma, and with his own particular environment. Thus, so far as observable characteristics are concerned, it frequently happens that the heterozygous individual, *Aa* or *aA*, appears superficially to be identical with one of the homozygous, *AA* say: in such a case *A* is said to be ' dominant ' and *a* ' recessive '. But, as has often been shown, what may appear identical to the unaided eye

abstract conception—' ein unbestimmtes Etwas mit ganz bestimmten Wirkungen; keine Hypothese über das Wesen dieses " Etwas " sollte dabei aufgestellt oder gestützt werden ' (28, p. 124). As a result of the work of Morgan and later cytologists, the word acquired a more specific connotation, viz., the minimum local unit in the material structure of the chromosome which is distinguishable from other units—distinguishable, that is to say, not only by its theoretical properties, but also as a result of actual breakage, crossing over, mutation, and the like. The present trend, however, is towards a re-interpretation of the gene as a functional rather than as a structural unit; and the familiar picture of discrete corpuscular entities, strung like beads on a linear thread, thus calls for appreciable modification, particularly in view of recent biochemical research. This will perhaps excuse what may seem to some a rather vague and loose use of the technical term gene in the discussion that follows.

Certainly, in the narrower sense of the term, there must unquestionably be far more genes than chromosomes. Hence it would be decidedly rash to assume that such genes must always be transmitted independently: those close together would tend to become linked, and would thus in certain cases be transmitted *en bloc*, as a kind of supergene (as indeed the evidence amply testifies). Moreover, when their effects were cumulative, the separate genes would no longer be identifiable. Hence for biometric purposes the genetic *factors* which we are here postulating may be regarded as often including groups of genes (in the more concrete and specific sense). On this account we think it wiser to describe our hypothesis as a ' multi-factorial ' hypothesis rather than as a ' polygenic ' hypothesis: see, however, Mather [32 and 34].

[1] The ordinary textbooks do not perhaps sufficiently insist that the epoch-making contribution of Mendel himself lay precisely in the point mentioned in the text, namely, in his recognition of the need for statistical treatment and of the consequent new for systematically planned experiments to verify the corollaries of his hypothesis—ideas which were fifty years in advance of those of his fellow biologists. Stripped of the numerical details, almost all the phenomena that he reports can be found, by those who have the curiosity to look, in previous articles in the *Journal of the Royal Horticultural Society*, notably those by Thomas Knight, who also hit on the culinary pea as a plant peculiarly suited for research. Galton's earliest experiments were on the inheritance of seed-weight in the sweet pea (*loc. cit.*, pp. 80 f.), and he to inclined to a 'particulate theory' (*loc. cit.*, pp. 7, 80f). In passing it is instructive to note that Galton and Mendel were both born in the same year (1822).

may be distinguishable by means of biochemical tests: ' dominance ', therefore, is in many ways a developmental or somatic feature, rather than a genetic feature.

An illustration will make this clear. On crossing reddish-purple peas with white, Mendel noted that in the first filial generation all the flowers were reddish-purple, while in the second both colours segregated in the proportion of 3 : 1. He therefore described the former colour as ' dominant '. The same phenomenon was observed when tall peas were crossed with short, and indeed in most of the characters studied by Mendel himself. On the other hand, when crossing red Antirrhinum (snap-dragon) with white, Punnett [49] obtained pink flowers in the first generation, and in the second white, pink, and red flowers in the proportions 1 : 2 : 1. Here the colour of the heterozygotes appeared to be an intermediate blend ; and other cases show that dominance may often be a matter of degree. In human beings, several specific forms of low grade intelligence (nearly all pathological) are now known to be the result of unifactor inheritance, either dominant, intermediate, or recessive (cf. Fraser Roberts [51] and Burt [3]). In what follows, however, we shall confine ourselves to normal (i.e., non-pathological) members of the population.

(*b*) *Multifactorial Inheritance.* When we turn to characters that appear to be the result of multifactorial inheritance, we have to suppose that the various genes concerned now all affect the *same* character, and are consequently responsible not for qualitative but for quantitative differences. The first to suggest this mode of inheritance was Mendel himself. When extending his experiments on hybridization from peas to beans, he discovered to his surprise that, on crossing *Phaseolus nanus* (which has white flowers) with *Ph. multiflorus* (which has purple) what he obtained was not one or the other of the original colours as with peas, but ' a whole series of colours from purple to pale violet and white ' (*eine ganze Reihe von Farben, von Purpur bis Blass-violett und Weiss*). After discussing the problem at some length, he concluded that ' auch diese rätselhaften Erscheinungen ' could be explained by the same ' laws ' as in the case of peas, provided we assume that the colours are due, not to a *single* 'factor ' (*differirende Merkmal*) *A* and its alternative *a*, but to a number of such factors, ' $A_1+A_2+...$ ' and their alternatives, none of which dominates the other and which together produce the *Gesammt-Eindruck der Farben:* in short, that the graded colours are the outcome of a 'Combinirung mehrerer selbständiger Farbenmerkmale '.[1]

Mendel's Model: Determination of Genotypes by Factor Analysis. The consequences will become obvious, if, following Mendel, we consider the case of two factors only. We begin (let us suppose) by crossing two true-breeding parental strains—red and white. With peas, where only a single factor is involved, the formulae for these homozygous parents will be *AA* and *aa*. Since in sexual reproduction *one* factor must come from each parent, all the first generation will necessarily have the heterozygous formula *Aa*; and their flowers will be red, because in peas that colour is dominant. On the other hand, with beans, where more than one factor is involved, the homozygous parents (we will suppose) have the formulae $A_1A_1A_2A_2$ and $a_1a_1a_2a_2$. From these therefore we shall obtain a first generation having in every case the formula $A_1a_1A_2a_2$. And, since complete dominance, is unlikely the flowers will be reddish, but not so dark as those of the red parent. We now cross these hybrids. What will be the formulae for members of the next generation?

The female gametes (the egg-cells) can carry only *one* of each kind of allelomorph. They therefore must have the formula A_1A_2, A_1a_2, a_1A_2, or a_1a_2 ; and these will, on an average, occur in equal numbers. The same four possibilities will arise in the case of the male gametes (the sperm- or pollen-cells) which fertilize them. But which particular type of sperm-cell fertilizes which particular type of egg-cell is in every case wholly dependent on chance.[2] What frequencies then are we likely to obtain?

The values to be expected in this case and in more complex cases may be readily deduced by methods similar to those of a simple factor-analysis.[3] Here a ' hierarchical matrix ' is constructed by pre-multiplying a row vector, representing the female gametes, by a column vector, representing the

[1] Mendel (35, pp. 29-34). These pages seem to have been overlooked by several writers of elementary textbooks, who state that ' this hypothesis, which represents one of the significant advances in genetic thought...was first proposed around 1910 ' (55, p. 314). A similar mistake seems to have been made by Punnett (49, p. 163).

[2] For class demonstrations the familiar method of setting students to draw counters from bags may be employed. To illustrate the argument above, we use black and white draughts (for the dominants and recessives), drawn by a ' father ' and a ' mother ' respectively, placing the dominant on top. The experiment is especially instructive because it brings home the kind of divergence from the theoretical ratios that is to be expected when the number of ' children ' is comparatively small.

[3] For the method of factor analysis, cf. Burt [4]. The device is particularly useful when the different kinds of gamete occur in unequal proportions (e.g., when the relative frequencies differ owing to linkage or the like, and we require to predict the relative frequencies of the F_2 genotypes to which they give rise). For other instructive applications of matrix methods (involving the determination of principal components and latent roots) see Fisher [22] esp. chaps. iii and iv, and Mather [34] chap. iv: also Hogben, *Proc. Roy. Soc. Edin.*, LIII, pp. 7-25.

male gametes ; and the products obtained will be those set out in Table I. We note that the four types of plant having a formula that appears in the principal diagonal will breed true, while the others, when self-fertilized, will in the next generation yield combinations whose formulae and frequencies can readily be ascertained by the same procedure as before. Thus, by raising yet another set of families we can obtain a simple test to clinch the accuracy of the theory.

TABLE I. MATRIX OF GENOTYPES FOR SECOND FILIAL GENERATION

The figures in bold type under each letter-group indicate the resulting grade or strength of the character

Male Gametes (Sperms)	Female Gametes (Eggs)			
	A_1A_2	A_1a_2	a_1A_2	a_1a_2
A_1A_2	$A_1A_1A_2A_2$ **4**	$A_1A_1A_2a_2$ **3**	$A_1a_1A_2A_2$ **3**	$A_1a_1A_2a_2$ **2**
A_1a_2	$A_1A_1A_2a_2$ **3**	$A_1A_1a_2a_2$ **2**	$A_1a_1A_2a_2$ **2**	$A_1a_1a_2a_2$ **1**
a_1A_2	$A_1a_1A_2A_2$ **3**	$A_1a_1A_2a_2$ **2**	$a_1a_1A_2A_2$ **2**	$a_1a_1A_2a_2$ **1**
a_1a_2	$A_1a_1A_2a_2$ **2**	$A_1a_1a_2a_2$ **1**	$a_1a_1A_2a_2$ **1**	$a_1a_1a_2a_2$ **0**

In accordance with the multifactorial hypothesis we here assume that every ' gene ' has similar effects, i.e., that its positive allelomorph (A_1 or A_2) will add a small unit of reddish colour to the petals (for instance, by attaching yet another hydroxyl group to a benzene ring in the pigment molecule), while the other allelomorph will add nothing. In the simpler cases dominance does not occur: so that *two* positive allelomorphs of the same type produce twice the effect of one. In such a case the maximum will be **4** units ; and we thus obtain five colour grades in all. The relative frequencies of each grade can readily be determined by counting up the number of instances in which there are **4, 3, 2, 1,** or **0** capital letters in the formula (as shown by the bolder figures under each letter-group). We obtain the frequencies shown in Table II. With a larger number of genes, n say, the number of grades will be $2n+1$, and the theoretical proportions of each can be found by expanding the binomial $(\frac{1}{2}+\frac{1}{2})^{2n}$.

TABLE II. FREQUENCY DISTRIBUTION OF PHENOTYPES

	Grade of Colour					Total
	4	**3**	**2**	**1**	**0**	
Theoretical Frequency	1	4	6	4	1	16

Evidence for Multiple Factors. Those psychologists who have discussed the multifactorial hypothesis still seem for the most part to suppose that the notion of multiple factors is a ' purely speculative conjecture advanced by over-zealous hereditarians to account for the continuous distributions so commonly found in human traits '. The earlier critics cited Johannsen's experiments on variations within pure lines of beans [28], and claimed that the results showed how ' all instances of continuous variation conforming with the normal curve must be of environmental origin '. More recent critics protest that the multifactor theory ' will explain practically anything, and yet is neither tested nor testable ' ([16], p. 732 ; cf. [25], p. 27). Before proceeding further, therefore, it may be well to summarize very briefly the evidence for now accepting multifactor inheritance as indeed a *vera causa*.

Owing to the low fertility of his hybrids, Mendel himself failed to secure sufficient offspring to confirm his explanation: it remained, as he said, *eine blosse Vermuthung*, and attracted little attention even after his work was rediscovered in 1900.

The first to report clear evidence in favour of such a hypothesis was the Swedish investigator, Hjalmar Nilsson-Ehle, in a series of three papers published in 1908-1910 [41, 42]. While studying the inheritance of seed-colour in red- and white-grained wheats, he found that with different strains the hybrid red varieties appeared to throw whites in different but fairly definite proportions: the palest gave ratios of approximately 3 red to 1 white; slightly darker varieties ratios of about 15 : 1; and others ratios of about 63 : 1. These figures suggested the operation of three factors of *gleichsinnige Wirkung*. If the student will construct an 8×8 table by the method used for Table I, and take gametes with the alternative threefold formulae of $A_1A_2A_3$, $A_1A_2a_3$, $A_1a_2A_3$, $A_1a_2a_3$, $a_1A_2A_3$, $a_1A_2a_3$, $a_1a_2A_3$ and $a_1a_2a_3$, he will obtain seven grades, with relative frequencies of 1, 6, 15, 20, 15, 6, and 1, i.e., 63 reddish plants (of varying shades) to 1 white. If, however, one factor (a_1 say) is assumed to remain constant, the recombinations of the other two will yeild 15 reds to 1 white (as in Table II); if two factors remain constant, the recombinations of the other will yield 3 reds to 1 white. By continuing the inbreeding to the third and later generations Nilsson-Ehle was able to verify the adequacy of this interpretation.[1]

In the cases with which we ourselves are concerned, the hypothetical factors will be far too numerous to be traced independently in this way. However, in 1913 East and Emerson obtained confirmation for a still more general form of the theory in the course of experiments on maize and the tobacco plant. Their method was to cross two distinct varieties, e.g., a tall Black Mexican sweet corn with the short Tom Thumb pop-corn, or a long-flowered tobacco-plant with a short-flowered strain. The results were in striking contrast with those reported by Mendel for tall and dwarf peas and beans. For example, the parent strains of *Nicotiana* had corollas measuring about 40 mm. and 94 mm. respectively. The first filial generation (F_1) produced flowers whose length was intermediate (about 65 mm.). Moreover, the variations of the individual plants (measured by the coefficient of variation) were comparatively slight and not significantly greater than those in either parental group; like them, they were in all probability due almost entirely to minor non-inheritable fluctuations. The next generation (F_2) was raised from average specimens, and showed a variability more than twice as great. The third generation (F_3) consisted of four groups of plants, raised from specimens deviating from the average by different amounts; and here the average flower length for each group was almost the same as that of its parent (i.e., there was little or no regression), while the variability of each group fell somewhere between the variabilities of the F_2 and F_1 generations —differing, in fact, according to the number of genes for which the F_2 parent was apparently homozygous. In this experiment, it will be seen, the individual differences are no longer qualitative, but purely quantitative—no longer discontinuous, but apparently continuous. Nevertheless, as East points out, to interpret them on Mendelian lines, no further assumptions are required beyond those previously established by experiments with qualitative and discontinuous characteristics—viz., the assumptions summed up in our six postulates above [14, 15].

East's theories were at first received with considerable scepticism. Indeed, at that time the genetical knowledge and the statistical techniques needed to cope with such problems were still in their infancy; and for a while the brilliant experimental work of Morgan diverted the interest of geneticists into more profitable channels. Consequently, apart from one or two contributions by Fisher, 'polygenetics', as it is sometimes called, remained almost entirely neglected for the next twenty years.[2] Quite recently, however, increasing attention has been devoted to polygenic research. It has been found that the factors responsible for quantitative variation—the so-called 'polygenes' —are sometimes linked with 'major genes' that can be used as markers [34, 50]. The check thus provided places the multifactorial hypothesis beyond all reasonable doubt. And it now seems abundantly demonstrated that many of the most valuable characteristics of animals and crops—size, weight, yield, and resistance to disease, and, indeed, nearly all those traits that exhibit continuous or

[1] Recent cytological work has shown that the commoner varieties of bread wheat (*Triticum vulgare*) are not diploid organisms, but hexaploid. The somatic cells contain, not 2×7 chromosomes (as do the kinds of grass from which the neolithic cultivators probably evolved it), but $2 \times 3 \times 7$ chromosomes—a set of 7 being technically known as a 'genome'. Thus the frequencies observed by Nilsson-Ehle were doubtless due, not (as was formerly supposed) to three 'genes', but to the partial homology between the three 'genomes' (cf. E. R. Sears, 'Cytology and Genetics of Wheat', *Advances in Genetics*, II, pp. 239 f. and refs.). Here therefore is a case in which the 'genetic factors' (as noted above) emerge primarily as statistical concepts.

[2] To this extent Lysenko's highly publicized attack on 'Morganism-Mendelism' may seem justified. What he and his associates failed to perceive was that, although it made little direct contribution to the improvement of stock and crops, nevertheless, by solving the simpler unifactorial problems, Morgan's experimental work laid the foundations for the statistical solution of multifactorial problems.

graded variation and what was formerly called ' blended heredity '—conform to the conditions of multifactorial inheritance [cf. 50, 54, and refs.]. We may therefore plausibly infer that, if variations in human intelligence are in any degree inheritable, then the *a priori* probability in favour of a multifactor hypothesis is decidedly high.

Number of Factors. How many independent factors are we to assume? From the studies by Morgan it is estimated that the diploid nucleus of the banana-fly (*Drosophila*), which has only four pairs of chromosomes, contains between 10,000 and 20,000 genes, and that man may have at least five times this number [54]. Man, however, like the potato, is usually said to possess as many as 48 chromosomes (23 pairs of autosomes and 2 sex chromosomes). In a paper submitted at the recent Congress on Human Genetics by Dr. Ford and Dr. Hamerton (of the Atomic Energy Research, Harwell) it was reported that the correct number should be 46 ; and this seems fully confirmed by the photomicrographs exhibited by Dr. J. Tijo (a Chinese investigator now working in Spain). In what follows we shall assume that the factors with which we ourselves are here concerned are purely autosomal (i.e., are not carried by the sex chromosomes). Hence, if the chromosomes were always recombined intact, that would imply 22 pairs of genes at most, and a maximum of 45 grades.

But in point of fact chromosomes are by no means unbreakable. And, if we can take the so-called ' chiasmata ', visible during the first meiotic division, as indicating the frequency with which segments of chromosomes cross over and so get interchanged, then it is conceivable that the maximum number of grades might possibly be raised to something like ten times the figure just given. Accordingly, we need have little hesitation in assuming that the number of ' multiple factors ', n say, is large enough for us to ignore quantities of the order $1/n$ [cf. 34].[1]

Additive Scales. When the character to be studied is produced by so large a number of genes, it will be impossible to trace the nature and influence of each separate one. But if, as we have assumed, the effects of all the genes are similar and cumulative, the resulting phenotype will display in different individuals different degrees of variation along the same dimension. Hence the first step towards determining the causes of such variations will be to examine their frequency distributions. The distributions themselves can be most readily studied and compared if we can specify their means, their variances, and their higher moments. But to do this we need a *metric scale.*

This is not a problem peculiar to psychologists. Whether he is dealing with psychological measurements or with physical, an essential preliminary that confronts every biometrician is, as Mather points out, ' the problem of the *scale* to be used in measuring or representing the character ' [11, pp. 78 f.]. For assessing ' intelligence ' a variety of procedures has been employed—order in a ranked group, the number of items completed in a given time, the time required to complete a given number of items, and, most frequently of all, the intelligence quotient or ' I.Q.'. We cannot, however, automatically assume that the unit which is most convenient for practical purposes is therefore adequate for the requirements of theoretical research. The chief difficulty arises from three possible types of interaction : (i) interaction between allelomorphs of the same gene ; (ii) interaction between different genes ; (iii) interaction between the genetic factors and the environmental. As we are concerned with *average* effects, what we need is ' some type of scale whereby the various agencies—both genes and non-heritable factors—have properties which are "simply additive" on the average ' (*ibid.*, p. 79). In any research, therefore, it will be essential to satisfy ourselves that the method of measurement adopted meets these preliminary requirements as far as possible.

A partial check is provided by the symmetry or lack of symmetry of the frequency distributions. With psychological tests depending wholly or partly on acquired skill or knowledge, there are often marked departures from the symmetrical form. Parents of high intellectual ability tend to create for their children an environment which favours intellectual acquirements, particularly of a verbal or literary type. Hence examinations in English or mathematics are apt to display a distribution with a tail elongated in the upper direction. The same holds good of tests of intelligence that depend largely on verbal ability (e.g., the Terman-Binet). When the tests are mainly or wholly non-verbal, and particularly when the test-results are checked and corrected by supplementary information, the divergence from normality tends to disappear ; indeed, with tests of the performance type, such as we ourselves have regularly used for borderline cases, it is hard to see how differences in the cultural environment could have any appreciable influence. On the other hand, forty years ago, when poverty was rife and infantile mortality high, there can be little doubt that malnutrition and disease, particularly during the first few years of life, might operate to enlarge the downward tail ; and a negative asymmetry of this kind was repeatedly discernible in the data from our earlier London

[1] The earlier investigators generally postulated a comparatively small number : East, for example, suggested 8. ' Student ', analysing results from a fairly thorough investigation (on oil content in maize), inferred that the number involved was at least 20 and probably nearer to 200. Later investigators have usually inclined towards this higher number. For methods and for criticisms see Woolf [50], pp. 96 f.

surveys. More recently, however, the distortion has become so slight that it can be demonstrated only in large samples, and then seems attributable chiefly to the effects of unifactor inheritance.[1]

Several writers have put forward methods for testing the adequacy of the scale to be used, or for transforming the scale in order to improve its adequacy ([34], pp. 40-51, [50], pp. 7-13); but, as ordinarily described, these are applicable only to experimental populations. With plants and animals it is generally possible to isolate two true-breeding lines, cross them, and then backcross the first generation of the cross (F_1) with one or other of the parental groups (P_a or P_b). If $\bar{B}_a$ denotes the mean for individuals in the backcross with parents whose mean is P_a, then, when the scale is additive, we ought to obtain $\bar{B}_a = \frac{1}{2}(\bar{P}_a + \bar{F}_1)$; and similarly $\bar{B}_b = \frac{1}{2}(\bar{P}_b + \bar{F}_1)$. For the next generation we should have $\bar{F}_2 = \frac{1}{2}(\bar{B}_a + \bar{B}_b)$, and so on. (For methods of applying these tests to actual data, see Mather [34], pp. 44-51.) With human beings deliberate experiments of this kind are out of the question. But the recent mingling of social classes and of rural populations previously isolated should make it possible to choose cases very roughly conforming with these requirements. The rather meagre results that we ourselves have collected along these lines appear consistent with the view that the I.Q. (especially when based on non-verbal tests) furnishes an additive scale, although little weight can be attached to so crude a check.

The method of checking the additive character of the inheritable and non-inheritable contributions is rather different. Before we can safely assume that the relevant environmental differences make their contribution to the total variation (or rather to the quantities that represent it) independently of the genetic differences, we must show that, on the scale proposed and within the margin of error allowed, the variances of the frequency-distributions for the separate lines and for their F_1's are the same. Here again our own results apparently conform with this requirement. In cases where the scale used for the raw measurements flagrantly conflicts with these conditions, better distributions may usually be secured by transforming the raw measurements –trying the logarithms, squares, or other powers, according to the familiar devices adopted by the psychologist when he wants to convert an asymmetrical distribution to a symmetrical.[2] When it proves impossible to eliminate non-additive interaction altogether, we may treat the total variance as the sum of the variances for the additive genetic and non-genetic components respectively, together with a joint residual term. With psychological measurements a further check is obtainable by correlating the assessments for the individuals with assessments of the relevant conditions in the environment. If the correlation is virtually nil, then we may reasonably infer that the genetic component is also uncorrelated with the environmental conditions. If however the correlation is significant, it does not *necessarily* follow that the genetic component is also correlated.

The Problem of Dominance. In considering the effects of a multiplicity of genes, we have to bear in mind that the allelomorphs of each dimorphic gene may (as we know from the study of qualitative characteristics) differ in three ways: (i) one allelomorph may be more favourable than the other, i.e., it may make a larger positive contribution to the quality in question or to its optimum amount; (ii) one allelomorph, whether favourable or not,[3] may show a greater degree of dominance than the other; and finally (iii) one allelomorph may occur more frequently than the other.

The phenomena of dominance create a special difficulty at the outset, particularly as dominance seems so often incomplete. In the case of quantitative characteristics the appearance of partial dominance is to some extent a matter of scaling. Thus with certain strains of pea (which Bateson termed 'dwarf' and 'half-dwarf') the average heights of homozygous and heterozygous specimens (or of what were assumed to be such) were found to differ approximately as follows (we have slightly modified the actual figures to make the comparison quite clear): *aa*, 0·7 m.; *aA*, 1·0 m.; and *AA*, 1·18 m. Here *A* appears to be both favourable and partially dominant, since the heterozygous form *aA* seems to resemble *AA*

[1] With plants and occasionally with animals, quantities like height, weight, and size (those which change with growth) frequently seem to yield skew distributions when measured in ordinary units. It has been suggested that this may arise from the fact that growth follows an exponential curve; and certainly a logarithmic scale often serves to restore normality [50, 19]. Asymmetry, as we shall see in a moment, may also be the effect of selection.

[2] In psychological work the most obvious case in which such transformations are needed arises with measurements involving short intervals of time: here, owing to the fact that zero time abruptly limits the scale in one direction, the initial distributions are apt to be asymmetrical. The detailed evidence for the conclusions reached in the text we hope to publish in a later communication on the general problem of transforming scales.

[3] Under natural conditions the incidence of dominance is by no means fortuitous. Dominance is produced by selection; and as a result unfavourable allelomorphs tend to become recessive and favourable to become dominant (cf. [19], pp. 48 f.).

(the tall homozygous form) much more than it resembles *aa* (the shorter form). If, however, we convert the measurements to logarithms, we obtain $aa = 5{\cdot}0$, $aA = 10{\cdot}0$, and $AA = 15{\cdot}0$. With this new scale the apparent dominance completely disappears, and the heterozygous form proves to be exactly intermediate, as if the two purer forms had blended. Of course when we are dealing with the effects of numerous genes, we cannot assume that precisely the same transformation would be equally successful with all of them. The obvious policy therefore will be to construct some hypothetical scale on which the genotypic measurements would be additive *on the average*.

The general principle can best be explained if we adopt a symbolic notation to express in algebraic form the two modes of measurement, phenotypic and genotypic. Consider first the case of a single gene, A-a say. Let $2a$ represent the difference between two individuals, reared in the same environment, and differing genetically only by being homozygous for opposite allelomorphs (or, if the data require it, let it represent the average difference between two such *groups* of individuals, reared under typical if varying environmental conditions). Then, when the *phenotype* for the heterozygous form appears to lie midway between the phenotypes for the two pure or homozygous forms, we may say that dominance is entirely absent; and when the former is identical with one of the latter, we may say that dominance is complete. These are limiting conditions. To cover the remaining cases let us measure the departure of the phenotype from the neutral midpoint by the quantity d. Then, unless $d = 0$, the increments or intervals from $-a$ to d and from d to $+a$ will *not* be equal.

However, since dominance is a purely somatic characteristic, we may suppose that the hypothetical intervals between the *genctypes aa* and *aA*, and again between *aA* and *AA*, are in fact really equal. Let us therefore denote their hypothetical measurements by the linear series $c-b$, c and $c+b$. We then require an equation of transformation which will enable us to replace the observed measurements $-a$, d, and $+a$ by the hypothetical quantities $c-b$, c, and $c+b$. In doing so we can at the same time allow for the fact that the observable differences between the *phenotypes* may not be, in any simple fashion, strictly additive. Here and elsewhere, when we require to circumvent the results of non-additive influences or interactions, the obvious procedure will be to invoke the method of least squares ; in other words, we shall seek those transformations which will minimize the error arising from the fact that we are substituting a linear function for a non-linear. In the present case, therefore, we shall choose c and b in such a way that the square-sum of the discrepancies between the genotypic values and the corresponding phenotypic values is a minimum : (the working formulae will be described below, p. 108). And this will yield the theoretical scale required.

The Ratio of Heterozygotes and Homozygotes. In experimental work, where the initial cross is between two true breeding and therefore purely homozygous lines, it is possible to ensure that the homozygotes will be present in equal numbers. Indeed, most of the formulae, given in earlier textbooks for correlations and the like, are based on this assumption. The student, however, should note that, in dealing with populations breeding freely, equality can no longer be validly assumed. As a rule, the dominant or ' wild ' form is likely to be the commoner, since in the wild state unfavourable dominants will be far more rapidly weeded out.

A simple mathematical deduction, however, leads to what is known as the Hardy-Weinberg law [23, 59], namely, that with random mating the three genotypes will, in a single generation, reach certain definite proportions (regardless of their initial proportions), and will thereafter remain constant, unless disturbed by mutations or differential selection. Thus, if P_1, P_2, P_3 denote the relative frequencies of *AA*, *Aa*, and *aa* respectively, then genetic stability will be attained when $P_2 = 2\sqrt{(P_1P_3)}$. In the case of human populations, there is ample evidence to justify the acceptance of this law as furnishing a reasonably close approximation. In what follows, therefore, we shall assume that the several proportions for *AA*, *Aa*, and *aa* are $P_1 = p^2$, $P_2 = 2pq$, and $P_3 = q^2$ respectively, where $p^2+2pq+q^2$, and therefore $p+q, = 1$.

The notation we have proposed for the various measurements and their frequency may accordingly be summarized by the following table.

TABLE III. NOTATION

Genotype	Phenotype	Phenotype measured from mean (m)	Contributions of additive portion	Differences due to dominance or interaction	Frequency
AA	a	$a-m=i$	$c+b$	$c+b-a$	p^2
Aa, *aA*	d	$d-m=j$	c	$c-d$	$2pq$
aa	$-a$	$-a-m=k$	$c-b$	$c-b+a$	q^2

II. THE UNIVARIATE DISTRIBUTION

The Curve of Distribution. Consider first a simplified model based on the following assumptions: (i) n independent loci; (ii) two allelomorphs for each locus; (iii) random determination of each resulting genotype; (iv) equal amounts added by every positive allelomorph; (v) strictly additive effects (no dominance or linkage). If the proportions (p_i) at different loci are not the same, the frequencies of the various genotypes would be given by $(p_1+q_1)^2(p_2+q_2)^2...(p_n+q_n)^2$: see Table I. But if n is large, we shall commit no serious error if we assume (vi) $p_1 = p_2 = ... = p_n = p$. Then the distribution of the phenotypes will be given by $(p+q)^{2n}$. Since with no dominance there will be three distinct phases for each gene, there will be 3^n different genotypes. But, as the value of each individual's measurement depends solely on the number of the positive allelomorphs, there will be only $2n+1$ grades: (cf. Tables I and II, p. 100).

If p differs widely from q, then, when n is small, the resulting distribution will be more or less skewed. But, as n increases, the asymmetry will tend to disappear: (to the unaided eye the frequency-polygon given by $(0{\cdot}9+0{\cdot}1)^{100}$, for instance, is almost indistinguishable from a continuous normal curve: see Yule and Kendall, [67], p. 173, fig. 10.1); and if there are any minor discontinuities, they will be completely smoothed away by the effects of environmental influences.

With human characteristics some degree of dominance seems far commoner than lack of dominance. If, then, one type of allelomorph (say the positive) always tended to show more dominance than the other, that would inevitably introduce a certain amount of asymmetry. But, except where some active form of selection is operative, no such one-sided association is discoverable; and, as Fisher has shown [19], the conditions of stability imply that the effects of dominance will in general be counterbalanced by the differences in the gene-ratios.

The criteria for normality (cf. Yule and Kendall [67], pp. 161 f.) are that the two beta-functions M_3^2/M_2^3 and M_4/M_2^2 should be 0 and 3 respectively, where M_2 is the second moment (σ^2), M_3 the third, and M_4 the fourth (all being obtained by summing the corresponding moments for the separate genes). Starting with expressions for the third and fourth moments analogous to those we used in calculating the variance for a single gene (eqn. (2) above), and collecting terms as before, we obtain (after a little elementary algebra) expressions for the two ratios showing that they depart from the requisite values only by quantities of the order $1/n$, and approximate more and more closely to those values as the number of genes is increased. Accordingly, an essential preliminary to investigating whether or not the multifactorial hypothesis is applicable to a given set of data will consist in testing the frequency-distribution for normality.[1] In the more exceptional cases in which selection is operative, the longer tail of the distribution will contain the phenotypes with the rarer and recessive allelomorphs, and will thus point away from the direction in which the selective agencies are working; possibly this may in part account for the slight asymmetry that occasionally appears (as noted above) in certain surveys made with intelligence tests.

Certainly with human measurements, both physical and mental, significant departures from the normal are at times observed. Often, as we have seen, they may be due to the metrical bias introduced by an inappropriate choice of scale. Still more frequently they may result from the disturbing effects of environmental influences. When they seem to be of genetic origin, we have still to determine their probable source—whether it be selection, an admixture of unifactor inheritance, a disturbance of the balance between effects of dominance and effects of gene-frequencies, or some other causal mechanism yet more elusive.[2]

III. THE COMPONENTS OF VARIATION

Total and Constituent Variations. The ultimate aim of our inquiry is to unravel the causes of individual differences in quantitative characters like height or intelligence. For specifying the amount of individual variation, the most convenient measure to adopt will be the variance, since variances by their very nature are both positive and additive. Apart from deliberate selection, the individuals whose characteristics we measure will (*a*) be drawn from different families or stocks and (*b*) be brought up in different environments. The

[1] Cf. Burt [7], p. 171 and refs.; and for the use of third degree statistics in this connection, see Fisher [20].

[2] Mr. Lewis [29] has recently criticized the alleged normality of many of the distributions reported on the ground that it is really an artefact resulting from the way the scale has been chosen. We therefore propose to examine this problem more fully in a later paper.

two most obvious sources of the observable differences therefore will be (1) heredity, H, and (2) environment, E.

Both words are here used in a rather broad sense: by heredity we mean all those tendencies which are transmitted by or through the reproductive mechanisms of the parents; and by environment, all the non-hereditary causes of individual variation, whether pre-natal or post-natal, intra-somatic or extra-somatic. Since we are here interested primarily in the operation of heredity, the statistical analysis will be greatly simplified if, in planning the preliminary collection of data, we endeavour to eliminate or at least to randomize the environmental variations. As this may not always prove possible, we may have to allow for a third component to represent the interaction or co-variance between the two.[1]

The Mendelian theory of inheritance implies, as we have seen, that the first of these components may itself be subdivided. It includes (1*a*) a portion which depends on the average difference between the two antithetical homozygous conditions arising from each gene-pair, and (1*b*) a portion which depends on the deviations of the heterozygous condition from the mean of the two homozygous conditions. Homozygotes breed true. Hence they and their progeny exhibit differences of the former type only. Consequently this kind of variation may be described as 'fixable': Fisher refers to it as 'variation due to the essential genotypes'. We shall therefore denote it by G. The other portion will depend in theory on the presence and amount of dominance. Clearly it is 'non-fixable'. Fisher sums it up as 'dominance deviations'. We shall therefore denote it by D. In human communities, particularly where monogamy prevails, conventional barriers greatly restrict the choice of a mate. Hence human parents tend to resemble one another in heritable characteristics. We shall therefore have to include (1*c*) a third component, M, to express the effects of assortative mating. There may occasionally be other sources of variation which ought strictly to be treated as genetic components—e.g., irremovable or unanalysed interactions between genes; if so, with the procedure here followed, they will probably become included with the non-heritable components; but their effects seem too small to be separately analysed, and in any case the error will be on the safe side.

For our purpose it will be convenient to think of the non-heritable component as also divisible into three parts: (2*a*) random environmental effects (E_r); (2*b*) systematic environmental effects (E_s); and (2*c*) a residual term obtained by replication (U), which, in an adequately planned inquiry, should be available for testing statistical significance. In the present investigation we shall not attempt anything like a rigorous analysis of these non-heritable variations; and, in the case of the physical measurements, we shall find that the last three components appear so small that they may be permissibly pooled.

Assuming that, in theory at least, all these components are additive, and allotting explicit weights (k) to indicate the relative influence of each, our analysis may be expressed as an equation:

$$V = k_h H + k_e E = k_g G + k_d D + k_m M + k_r E_r + k_s E_s + k_u U, \quad \text{or} \quad k_g G + k_d D + k_e E,$$

if mating is random and the non-heritable variance pooled.[2] The analysis itself will be

[1] Referring to a previous paper by Burt, Dr. Barnet Woolf 'wonders what meaning, if any, these authors attach to the term "environment"', ([50], p. 91). He has evidently failed to note the discussions of 'environmental conditions' in previous publications: see below, p. 127.

For some reason that is not made clear he also holds that, in investigations like the present, it is a form of 'legerdemain' to attempt to 'eliminate' the effects of environment; they should, he contends, be preserved 'as an integral part of the situation'. But surely, when investigating the effects of genetic conditions, it is a sound rule to minimize, so far as possible, the irrelevant influence of environment (e.g., by an appropriate selection of tests or of persons), just as, when investigating the effects of different environmental conditions, it is helpful to minimize the varying influence of heredity. In regard to Dr. Woolf's more general suggestion that, in these and analogous investigations, first degree statistics should be substituted for the analysis of variance, the reader may be referred to Dr. Mather's reply (*loc. cit.*, pp. 103-111). We hope to deal with these and other objections more fully in the next issue of this *Journal*.

[2] To a large extent the analysis is the same as that developed by Sewall Wright, Mather, Lush and others to deal with quantitative inheritance in plants and animals. It has recently been subjected to a somewhat unconvincing criticism by Woolf ([50], pp. 81-102); Mather has contributed a reply in the same publication (pp. 103-113).

based on second order statistics derived from the data, namely, variances and covariances or coefficients of correlation. Correlations therefore we shall regard as expressing, not so much an alleged resemblance, but rather a relation between certain variances.

Criticisms of this and similar schemes can best be dealt with after seeing how they work out in actual practice. But one minor point should perhaps be emphasized forthwith. A variance is essentially an average. Hence our conclusions hold only of groups, not of individuals. Critics like Dr. Heim [24] and Mr. Maddox [31], who object that *individual* measurements cannot be analysed along the lines proposed, overlook the fact that inferences based on an analysis of variance do not pretend to be applicable to anything but the *group* concerned, or the population of which the group forms a representative sample.

Outline of Argument. In deriving the requisite formulae the method we shall follow will be to start with the simplest type of situation and proceed step by step to the more complex. Thus we shall begin with formulae describing the results obtained in the case of a single gene (assuming two allelomorphs only), and from these we shall build up the results to be expected with a number of such genes (assuming throughout that their effects are independent, i.e., that there is no linkage and no interaction). Moreover, in the first stages of the argument we shall also assume that mating is perfectly random. With this simplified scheme we can readily derive values for the parental, fraternal, and other correlations ; and we shall find that on the whole they agree very roughly, though by no means precisely, with the coefficients commonly reported. In order to improve the estimates so obtained we shall then go on to modify our simplified formulae by attempting to allow for the effects of assortative mating.

The final verification of our basic hypothesis will proceed as follows. We shall find that the assumptions we have made require the introduction of three essential parameters. We may estimate these either (*a*) from *a priori* arguments or (*b*) from empirical data (viz., some of the observed correlations). This will enable us to deduce theoretical values for those correlations which have not been needed for the estimation. And, if the observed values agree with the theoretical, we may then maintain that our hypothesis is confirmed.

The formulae employed will in the main be those deduced by Fisher. But to simplify the exposition we shall occasionally sacrifice rigour to brevity, or even offer an alternative proof. In certain respects too we shall introduce minor modifications that seem to be needed in adapting the general procedure to the peculiarities of psychological characteristics. But we hope that the ultimate result will be to send the interested reader to a first-hand study of Fisher's highly original investigations [17, 19, 20]. To facilitate a comparison with his demonstrations we shall here for the most part retain his notation.

IV. THE RESULTS OF RANDOM MATING

So long as we can assume that mating is entirely random, the deduction of the requisite formulae is perfectly straightforward, though sometimes a little lengthy. In the end, as we shall see, it leads to expressions which are surprisingly simple. It may be noted, however, that the adoption of a matrix notation (together with a few special symbols for certain recurrent operations) would greatly facilitate the construction of bivariate frequency-tables and the derivation of formulae for correlations between relatives, particularly when the relationships are somewhat remote. See below, Appendix, p. 131.

Formulae for the Total Heritable Component. Let us begin by considering a population in which the measurements for the particular trait in which we are interested are due solely to a single dimorphic gene A-a. On multiplying the measurements for the three phenotypes by their proportional frequencies we at once obtain

(1) for the mean,

$$m = p^2a+2pqd-q^2a,$$

so that

$$p^2(a-m)+2pq(d-m)-q^2(a+m) = 0. \qquad (1)$$

Similarly we obtain
(2) for the variance,

$$\begin{aligned}\alpha^2 &= p^2(a-m)^2+2pq(d-m)^2+q^2(a+m)^2 \\ &= (p^2+q^2)a^2+2pqd^2-2(p^2a+2pqd-q^2a)m+m^2 \\ &= \{(p^2+q^2)-(p^2-q^2)^2\}a^2+4pq(p^2-q^2)ad+2pq(p^2+q^2)d^2.\end{aligned} \tag{2}$$

But $p+q=1$, and therefore $p^2+q^2=1-2pq$ and $(p^2-q^2)^2=(p+q)^2(p-q)^2=(p-q)^2$. Hence

$$\alpha^2 = 2pq[(a^2-2(p^2-q^2)ad+(p^2+q^2)d^2] \tag{3}$$

By the terms of our hypothesis the gene A-a is only one of a very large number, n. If in the general population these different genes are associated at random, then the total variance resulting from their cumulative effects will be $\sum\alpha^2=\sigma^2$ (in Fisher's notation) $=H$ (in the notation suggested above).

Fixable and Non-fixable Components. We have now to allow for the fact that there may be departures from the simple addition of the separate effects of each individual gene, of which the most important are those due to the dominance of one or the other of the paired allelomorphs. They may be treated statistically along the lines suggested in our discussion of scaling. By applying the method of least squares we can analyse the contributions of imperfectly additive genes into (i) a hypothetical component which will represent their genetic effects with as little distortion as possible, and (ii) a residual portion which will operate rather like an arbitrary error obscuring the undistorted measurements.

As before, let us consider first the contributions of a single gene in its three different phases. And, in accordance with the notation already proposed, let $-a$, d, and $+a$ denote the observable results. For these let us now substitute (as suggested above) the linear values $c-b$, c, and $c+b$, choosing values for c and b such that the weighted sum of the squared discrepancies

$$p^2(c+b-a)^2+2pq(c-d)^2+q^2(c-b+a)^2=\delta^2$$

say, is a minimum. On differentiating this expression with respect first to b and then to c, and setting the resulting expressions equal to zero, and finally dividing all through by 2, we obtain

$$(p^2+q^2)b+(p^2-q^2)c-(p^2+q^2)a=0,$$

and

$$(p^2-q^2)b+c-(p^2-q^2)a-2pqd=0.$$

Solving for b and c, we find

$$b=a-\frac{2pq(p^2-q^2)d}{(p^2+q^2)-(p^2-q^2)^2}=a-(p^2-q^2)d, \tag{4}$$

and

$$c=\frac{2pq(p^2+q^2)d}{(p^2+q^2)-(p^2-q^2)^2}=(p^2+q^2)d, \tag{5}$$

since, as we have already seen, $(p^2+q^2)-(p^2-q^2)^2=2pq$.

We now insert these values in the expression just given for the weighted sum of the squares. Noting that $p^2+q^2-1=p^2+q^2-(p^2+2pq+q^2)=-2pq$, we obtain

$$\begin{aligned}\delta^2 &= p^2(2q^2d)^2+2pq(p^2+q^2-1)^2d^2+q^2(2p^2d)^2 \\ &= 4p^2q^2(q^2+2pq+p^2)d^2=4p^2q^2d^2=\tfrac{1}{4}d^2,\end{aligned} \tag{6}$$

if $p=q=\frac{1}{2}$. This quantity then represents the contribution to the variance of the irregular or residual portion.

Now let β^2 denote the contribution of the strictly additive portion. Then

$$\begin{aligned}\beta^2 &= p^2(c+b-m)^2+2pq(c-m)^2+q^2(c-b-m)^2 \\ &= (p^2+2pq+r^2)(c-m)^2+(p^2+q^2)b^2+2b(p^2-q^2)(c-m) \\ &= (c-m)\{(c-m)+2b(p^2-q^2)\}+(p^2+q^2)b^2.\end{aligned}$$

But from eqn. (1) we have for the mean

$$p^2(c+b-m)+2pq(c-m)+q^2(c-b-m)=0,$$

i.e., $(c-m)+b(p^2-q^2)=0$, and therefore $c-m=-b(p^2-q^2)$.

Hence

$$\beta^2=\{(p^2+q^2)-(p^2-q^2)^2\}b^2=2pqb^2=2pqb^2=\tfrac{1}{2}b^2, \tag{7}$$

if $p=q=\frac{1}{2}$.

Again, on substituting in eqn. (3) the value for a given by eqn. (4), we have

$$\alpha^2=2pq\{b^2+[(p^2+q^2)-(p^2-q^2)^2]d^2\}=2pqb^2+4p^2q^2d^2 \tag{8}$$

$$=\beta^2+\delta^2. \tag{9}$$

The argument may be generalized to cover each and all the various genes whose effects we have to consider. If, therefore, we write $\sum\alpha^2 = \sigma^2$, $\sum\beta^2 = \tau^2$, and $\epsilon^2 = \sum\delta^2$, we have

$$\sigma^2 = \tau^2 + \epsilon^2. \tag{10}$$

We have thus analysed the total variance into two independent and contributory portions as required, namely, a strictly additive portion representing the fixable elements in the genetic component and a residual non-fixable portion resulting from the influence of dominance and the like. On putting $G = \sum b^2$ and $D = \sum d^2$ in eqn. (8), we obtain, for the special case in which $p = q = \frac{1}{2}$, the unexpectedly simple equation,

$$H = \tfrac{1}{2}G + \tfrac{1}{4}D. \tag{11}$$

V. BIVARIATE DISTRIBUTIONS

Correlations between Relatives. To obtain formulae for the parental, fraternal, and other correlations, we may proceed as follows. Beginning once again with a single gene, let us express the observed phenotypic effects in the form of deviations from the mean (see Table III). Two constants only will be necessary, since from eqn. (1) we have $p^2i + 2pqj + q^2k = 0$, and therefore

$$j = -\frac{p^2i + q^2k}{2pq}. \tag{1 \textit{bis}}$$

For the moment we have agreed to assume that mating is entirely random. We may therefore think of marriage and its results as a perfect lottery, in which balls of different colours mixed in specified proportions are drawn blindly from an enormous bag. On this basis, by invoking the elementary principles of probability we can readily construct 3×3 distribution tables showing the expected correlations between relatives of various degrees of affinity.

Possible Unions. Consider first the relative frequency of the different types of marital union that are possible. This is shown in Table IV. Under the conditions of random mating the expected proportions will evidently be given by a 3×3 'hierarchical' matrix, obtained by pre-multiplying the row of frequencies for wives by the column of frequencies for husbands.

TABLE IV. FREQUENCIES OF POSSIBLE UNIONS ARISING WITH RANDOM MATING

Husbands		Wives		
Type	*Frequency*	*Frequency* / *Type*	*AA* p^2	*Aa* $2pq$	*aa* q^2
AA	p^2		p^4	$2p^3q$	p^2q^2
Aa	$2pq$		$2p^3q$	$4p^2q^2$	$2pq^3$
aa	q^2		p^2q^2	$2pq^3$	q^4

A. *The Parental Correlation.* To obtain the correlation between parents and children let us suppose that each family is of the same size, and then pair each child with its father. Take first the p^2 fathers belonging to the type AA. From an AA father and an AA mother, nothing but AA children will be born; and of these (as Table IV shows) there will be $p^2 \times p^2 = p^4$. From an AA father and an aa mother, nothing but Aa children can be born; and of these there will be $p^2 \times q^2 = p^2q^2$. From AA fathers and Aa mothers, the children born may be either AA or Aa, and we may expect equal proportions of each, viz., $p^2 \times pq = p^3q$. Thus the total number of AA children will be $p^4 + p^3q = p^3(p+q) = p^3$; and the total number of Aa children will be $p^2q^2 + p^3q = p^2q(p+q) = p^2q$. There can be no aa children born to an AA father. The proportions thus derived are set out in the top row of Table V. The remaining proportions can be deduced in the

same way.[1] The total proportions for the three types of child (shown at the foot of the table) are the same as the proportions for the three types of parent. Thus the population is genetically stable.

The product sum may be computed in the usual way, and yields the quadratic

$$p^3i^2+2p^2qij+2pq^2jk+pq(p+q)j^2+q^3k^2.$$

Substituting for j from eqn. (1 *bis*), we find that this reduces to $\frac{1}{4pq}(p^2i-q^2k)^2 = \frac{1}{2}\beta^2$. The variance is given by eqn. (9). Hence the product-moment correlation will be $\frac{1}{2}\left\{\frac{\beta^2}{\beta^2+\delta^2}\right\}$. When there is

TABLE V. CORRELATION BETWEEN PARENT AND CHILD

Parent	Child			Total
	AA(i)	Aa(j)	aa(k)	
AA(i)	p^3	p^2q	—	p^2
Aa(j)	p^2q	pq	pq^2	$2pq$
aa(k)	—	pq^2	q^3	q^2
Total	p^2	$2pq$	q^2	1

a large number of genes, the covariances and variances arising from each one may be summed as before.[2] Thus, on the assumption of multifactorial inheritance with random mating, the theoretical value for the parental correlation will be

$$r_{po} = \frac{1}{2}\left\{\frac{\tau^2}{\tau^2+\epsilon^2}\right\} = \frac{1}{2}\frac{\tau^2}{\sigma^2}. \tag{12}$$

If there is no dominance whatever, i.e., if the value for the heterozygote falls in every case midway between that of the two homozygotes, then $d = 0$, and consequently $\epsilon^2 = 0$, and $r_{po} = 0{\cdot}50$, a value which would agree reasonably well with the correlations actually obtained both for height and for intelligence. The effect of dominance, it will be seen, is to reduce the correlation in the ratio $\frac{\tau^2}{\sigma^2} = g^2$ (say). Whether dominance is present or not, the covariance between parent and offspring, as will now be clear, depends solely on what has been termed the *fixable* part of the genetic component. Thus the assumption, made by so many psychological writers, that the correlation measures the *whole* of the genetic contribution, is manifestly erroneous.

In experiments where the initial parents are deliberately selected from pure lines, and where, therefore, the gene-ratios are presumably equal (i.e., $p = q = \frac{1}{2}$), the parental correlation may be written

$$r_{po} = \frac{1}{2}\cdot\frac{\frac{1}{2}G}{\frac{1}{2}G+\frac{1}{4}D} = \frac{1}{2}g^2. \tag{12a}$$

In the earlier discussions of the Mendelian hypothesis, it was generally assumed, not only that the frequencies of the two pure phases were equal, but also that dominance was complete, i.e., $d = a$ and therefore $D = G$. With these assumptions we at once obtain $r_{po} = \frac{1}{2}\cdot\frac{2}{3} = 0{\cdot}3\dot{3}$, which is the value reached by Pearson.

The above formulae presuppose that non-heritable contributions are negligible. If they are large enough to be taken explicitly into account, we may write $\frac{\frac{1}{2}G+\frac{1}{4}D}{\frac{1}{2}G+\frac{1}{4}D+E} = h^2$ (say) for the

[1] For a more detailed algebraic deduction of the various frequencies, see David [12], pp. 90 f. If we assume that the proportions of the pure types are approximately equal, and put $p^2 = q^2 = 25$ per cent., we obtain the theoretical frequencies given in Table II of Burt's recent article ([7], p. 173): this answers the inquiry of several critics as to how the figures there set out were actually obtained.

[2] The justification for this procedure may be expressed as follows. Take the mean of the offspring in any one array of the bivariate table and consider its regression on the general mean. The regression due to a *single* gene will evidently be $\frac{1}{2}\beta^2/\sigma^2$. Summing the effects of *all* the genes we find that the resulting regression will be $\frac{1}{2}\Sigma\beta^2/\sigma^2 = \frac{1}{2}\tau^2/\sigma^2$. Since the standard deviations of both parents and offspring are equal, the regression will be identical with the correlation.

proportional amount contributed by heredity to the total variance. And the parental correlation then becomes

$$r_{po} = \frac{\frac{1}{2}G}{\frac{1}{2}G+\frac{1}{4}D+E} = \frac{1}{2}h^2g^2. \qquad (12b)$$

Of these two expressions, the former is that used by Mather in his experimental work [34, p. 78]; and the latter and more general version is that used by Sewall Wright [50, p. 17].

Wright has extended Fisher's analysis so as to allow for possible effects of interaction between genes [64, 65]. Thus, if i^2 denotes the variance due to interaction (expressed as a fraction of the total genetic variance), then, applying the same least squares procedure as before, we obtain

$$r_{po} = \tfrac{1}{2}h^2g^2+\left(\frac{p}{1+p}\right)^2 h^2i^2. \qquad (12c)$$

This version of the formula suggests that, even when present, the effects of interaction are not likely to produce any great alteration in the theoretical value of the correlation; and for the characteristics

TABLE VI. FREQUENCIES OF POSSIBLE SIBLINGS OCCURRING WITH RANDOM MATING

Frequency	Parents	Children	Pairs of Siblings						Total
			AA & AA	AA & Aa	AA & aa	Aa & Aa	Aa & aa	aa & aa	
p^4	AA×AA	All AA	1	0	0	0	0	0	1
$4p^3q$	AA×Aa	$\frac{1}{2}$AA+$\frac{1}{2}$Aa	$\frac{1}{4}$	$\frac{1}{2}$	0	$\frac{1}{4}$	0	0	1
$2p^2q^2$	AA×aa	All Aa	0	0	0	1	0	0	1
$4p^2q^2$	Aa×Aa	$\frac{1}{4}$AA+$\frac{1}{2}$Aa+$\frac{1}{4}$aa	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{16}$	1
$4pq^3$	Aa×aa	$\frac{1}{2}$Aa+$\frac{1}{2}$aa	0	0	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1
q^4	aa×aa	All aa	0	0	0	0	0	1	1

TABLE VII. CORRELATION BETWEEN SIBLINGS

Siblings	Siblings			Total
	AA (i)	Aa (j)	aa (k)	
AA (i)	$p^2(p+\frac{1}{2}q)^2$	$p^2q(p+\frac{1}{2}q)$	$\frac{1}{4}p^2q^2$	p^2
Aa (j)	$p^2q(p+\frac{1}{2}q)$	$pq(p^2+3pq+q^2)$	$pq^2(\frac{1}{2}p+q)$	$2pq$
aa (k)	$\frac{1}{4}p^2q^2$	$pq^2(\frac{1}{2}p+q)$	$q^2(\frac{1}{2}p+q)^2$	q^2
Total	p^2	$2pq$	q^2	1

with which we are here concerned no concrete evidence is available to suggest that any such interaction exists, much less to enable us to compute its amount.

B. *The Fraternal Correlation.* The determination of the covariances with other members of the family is rather more complicated. The student can most readily compute the relative frequencies if he tabulates the several possibilities at each successive stage in accordance with the scheme shown in Table VI for siblings. The frequencies for the several unions (first column) are taken from Table IV. They are, as we have seen, the values obtained by expanding the expression $(p^2+2pq+q^2)^2$. The frequencies for the several types of child (third column) are calculated much as before by considering the relative frequency with which one gene from the father will combine with one gene from the mother, when the combinations are assumed to be random. The frequencies for the various pairs of siblings are the numerical coefficients obtained by squaring the expressions shown in column 3 of Table VI. And we then multiply these last sets of frequencies by the proportional frequency shown at the beginning of the row (i.e., in column 1). The totals obtained in this way from each of the six columns of pairs have been entered in Table VII, all except those for the diagonal entries being split into equal halves, since a pair like AA and Aa in Table VI also includes the pair Aa and AA in Table VII. (See also Appendix, p. 131).

To determine the product-sum it is simplest to regard the frequency matrix thus obtained as the sum of two component matrices, one of which is identical with the frequency-matrix for parent and child. The product-sum will then consist of the same quadratic as was obtained for the parental correlation with the addition of a supplementary quadratic, namely,

$$\tfrac{1}{4}p^2q^2(i^2-2ij+2ik-2jk+4j^2+k^2) = \tfrac{1}{4}\delta^2.$$

Thus the total product-sum is $\frac{1}{2}\beta^2+\frac{1}{4}\delta^2$; and the product-moment correlation will be $\frac{1}{2}\left(\frac{\beta^2+\frac{1}{2}\delta^2}{\beta^2+\delta^2}\right)$. Summing for the effects of the various genes, we accordingly obtain for the fraternal correlation resulting from multifactorial inheritance

$$r_{oo'} = \frac{1}{2}\cdot\frac{\tau^2+\frac{1}{2}\epsilon^2}{\sigma^2}. \tag{13}$$

Evidently when there is no dominance, $r_{oo'} = \cdot 50$, the same value as that obtained under similar conditions for the parental correlation. Since $\sigma^2 = \tau^2+\epsilon^2$, the effect of dominance will be to reduce the fraternal correlation by half the amount by which the parental correlation is reduced.

This halving of the effect of dominance, it will be observed, has been deduced under very general conditions: it is independent of any variation in (*a*) the influence of the different genetic factors, (*b*) the proportions in which their different phases occur, and (*c*) the degrees of dominance. Provided the samples are suitably selected and sufficiently large, and the difference between the two observed correlations is statistically significant, this difference in the two formulae, as Fisher points out, should ' allow us to distinguish between the random external effects of environment and the effects of dominance '. However, if in the sample examined there is reason to believe that the effects of environment are not entirely random, i.e., if the character measured is correlated to an unknown extent with the environmental conditions, then the increase in the correlation between the siblings might conceivably be due to the similarity of their environment.

When $p = q = \frac{1}{2}$, the formula becomes

$$r_{oo'} = \frac{1}{2}\cdot\frac{\frac{1}{2}G+\frac{1}{8}D}{\frac{1}{2}G+\frac{1}{4}D} = \frac{1}{4}(g^2+1). \tag{13*a*}$$

Evidently, when dominance is complete, i.e., when $D = G$, $r_{oo'} = \frac{1}{2}\cdot\frac{5}{6} = \cdot 41\dot{6}$. This in fact is the value derived by Karl Pearson, and, like his estimate for the parental correlation, is far too low to fit the coefficients actually reported. If we take the random effects of environment into account, we have

$$r_{oo'} = \frac{\frac{1}{4}G+\frac{1}{16}D}{\frac{1}{2}G+\frac{1}{4}D+E} = \frac{1}{4}h^2(g^2+1). \tag{13*b*}$$

And if we remove the limitation on the values of p and q, and introduce a further component for interaction, this becomes

$$r_{oo'} = \frac{1}{2}h^2g^2+\frac{1}{4}h^2d^2+\frac{1}{16}\left(\frac{1+3p}{1+p}\right)^2h^2i^2 \tag{13*c*}$$

where d^2 and i^2 denote the variances due to dominance and interaction respectively, expressed as ratios of the total genetic variance. Once again we note that the additional term is not likely to produce any great alteration. However, since in practice the various types of non-additive influence are difficult to distinguish, ' it is preferable ', as Mather points out, ' to have the non-additive effects represented as dominance rather than as interaction ', and this can be accomplished by choosing a suitable scale [34, p. 141].

Analogous formulae for the correlations between remoter relatives can be deduced by the same procedure. The avuncular correlation, for example (the correlation between uncle or aunt and nephew or niece) proves to be exactly one half the parental, and the correlation between first cousins exactly one quarter the parental. There is one important correlation which is of special interest to the psychologist, the correlation between the *offspring* of first cousins. The requisite formulae can be derived by the procedures used above to determine the correlation for the offspring of unrelated parents. It is unnecessary to set out in full the complex expression that results; but if, as before, we take the simplest case (dominance absent and no sex-linkage, the correlation for siblings is raised from 1/2 to 9/17, i.e., from ·50 to ·53 and for parent-and-child to 6/11, i.e. ·55.

One incidental consequence of the argument employed above is worth noting. Under the conditions assumed, it will be seen that there is a suggestive analogy between the bivariate frequency table for a *single* gene and the corresponding table which would be obtained from summing the frequencies for *all* the n genes. Thus, if we take average figures for i, j, k, p and q, then theoretical estimates for the correlations may be readily deduced by dividing the population into three classes, having large, intermediate, and small measurements respectively for the quantitative trait: the limits

between the classes will be so chosen as to give the requisite average values for l, j, k, p and q. Where, for example, dominance is completely absent, or where it operates equally in both directions (which in practice will mean cases where its overall effects are so slight as to be non-significant with the numbers available), we can put $i, j, k = +1, 0$, and -1, and $p = q = 50$ per cent. The theoretical values for the bivariate distributions shown in Tables II and III of the article already cited ([7], p. 173) were obtained in this way.

VI. THE EFFECTS OF ASSORTATIVE MATING

We shall now remove the restriction that the mating of parents is always random. In man, for nearly all the commoner traits hitherto studied, there is ample evidence for a more or less two-sided sexual selection leading to positive correlations between husband and wife. The correlations are seldom high, yet often large enough to require appreciable changes in the formulae we have so far deduced.

(*a*) *With a Single Factor.* The general tendency of such preferential mating may be most readily understood if we glance first of all at the effects produced when only a single pair of genes is involved, and consider, for the sake of simplicity, the extreme and imaginary case in which every husband chooses a wife of his own genetic type. In such a case, instead of the nine values set out in the 3×3 table for random matings (Table IV), the frequency-matrix for husbands and wives would have the values p^2, $2pq$, and q^2 in the principal diagonal and zeros elsewhere. To determine the correlation between siblings, we should have to substitute these proportions for those shown in the first column of Table VI; lines 2, 4, and 5 would be deleted; and then, taking the fractions given in the remaining rows, we should obtain the bivariate distribution shown in Table VIII.

TABLE VIII. CORRELATIONS BETWEEN SIBLINGS

With perfect correlation between fathers and mothers

Siblings	Siblings			Total
	AA	Aa	aa	
AA	$p^2+\frac{1}{8}pq$	$\frac{1}{4}pq$	$\frac{1}{8}pq$	$p^2+\frac{1}{2}pq$
Aa	$\frac{1}{4}pq$	$\frac{1}{2}pq$	$\frac{1}{4}pq$	pq
aa	$\frac{1}{8}pq$	$\frac{1}{4}pq$	$q^2+\frac{1}{8}pq$	$q^2+\frac{1}{2}pq$
Total	$p^2+\frac{1}{2}pq$	pq	$q^2+\frac{1}{2}pq$	$p^2+2pq+q^2$

We note that the variance is increased, and that the population is no longer in stable equilibrium. If we assume that there is no dominance, the correlation will evidently be $r_{00'} = \frac{p^2+q^2}{p^2+q^2+pq}$. Thus when $p = q = \frac{1}{2}$, the presence of perfect assortative mating would raise the correlation from 0·500 to 0·66̇6̇. When the degree of assortative mating is incomplete, the fraternal correlation will lie somewhere between these two values, according to the amount of correlation between the fathers and the mothers.

(*b*) *With Multiple Factors.* With multi-factor inheritance the influence of assortative mating becomes far more complex. We now have to allow for the fact that there will be some degree of association between similar phases of *different* factors. To deduce the modifications thus entailed in the formulae previously derived involves a somewhat lengthy and elaborate statistical argument.[1] Here we can do no more than briefly indicate the main steps in Fisher's ingenious proofs.

Let μ denote the correlation between fathers and mothers, and let us assume that the population has reached a stable equilibrium. Let P, Q, and R represent the new proportional frequencies for the three genotypic phases, and let I, J, and K denote the means of the distributions of these three phases. As will be seen from Table VI, column 3, the instability we have noted arises from

[1] An alternative approach has been recently put forward by Sewall Wright, who suggests applying his method of 'path coefficients'. A convenient account of the procedure, with illustrated examples, will be found in E. C. R. Reeve and F. W. Robertson, 'Studies in Quantitative Inheritance, II. Analysis of a Strain of *Drosophila Melanogaster* selected for Long Wings', *J. Genetics*, LI, 1953, pp. 276-316.

marriages (i) between the two pure types (AA and aa) which yield offspring of a mixed type (Aa) and (ii) between two mixed types (Aa and Aa) which yield offspring half of whom revert to a pure type (AA or aa). For equilibrium the latter type of mating must be twice as frequent as the former. This yields the condition

$$PR-Q^2 = Q^2\mu\frac{J^2-IK}{V}+\ldots \tag{14}$$

where the terms omitted are squares of small fractions which can be neglected. If $\mu = 0$, this implies $PR = Q^2$, which, it will be remembered, was the condition of stability with random mating.

If, as before, the relative frequencies of pairs of allelomorphs are p and q, we obtain for the frequencies of the three phases $P = p^2+p^2q^2\mu h$, $Q = pq-p^2q^2\mu h$, and $R = q^2+p^2q^2\mu h$, where $h = (J^2-IK)/V$. Evidently the mean $PI+2QJ+RK = 0$. But it will now be seen that the 3×3 possible combinations of two genetic factors will no longer occur in the simple proportions, PP', $2PQ'$, etc., as in the case of random mating, but will be changed to (say) $PP'(1+f_{11})$, $2PQ'(1+f_{12})$, etc. It can then be shown that $I = i+\sum(P'if_{11}+2Q'f_{12}+R'kf_{12})$, where the summation extends over all the factors except those to which i, j, k refer. There will be analogous values for J and K. We thus eventually reach for the modified variance the equation

$$V = (PiI+2QjJ+RkK). \tag{15}$$

This implicitly involves the unknown quantities introduced above. Values for these, however, may be deduced by again invoking the conditions required when the population is in stable equilibrium. The final outcome is a comparatively simple expression for the variance which can readily be computed from the data we have postulated, namely,

$$V = \sigma^2+\frac{A}{1-A}\tau^2, \tag{16}$$

where

$$A = \frac{\tau^2\mu}{\sigma^2-A\epsilon^2} = \frac{\tau^2+\dfrac{A}{1-A}\tau^2}{\sigma^2+\dfrac{A}{1-A}\tau^2}\mu \tag{17}$$

$$= c_2\mu \text{ (say).} \tag{18}$$

Here c_2 represents the ratio of (i) the variance *without* the deviations due to dominance to (ii) the variance *with* those deviations. Thus, whereas μ denotes the crude value for the marital correlation based on the observed measurements, A is in effect an estimate of the marital correlation between the hypothetical values of the genotypes unobscured by the effects of dominance. It is evidently a quantity varying between 0 and μ. When dominance is entirely absent, $A = \mu$; when mating is random, $A = 0$, and the equation for V (eqn. (16)) reduces to $V = \sigma^2$.

Before going on to determine the modified values for the correlations between relatives, we must first re-examine the effects of environment and of dominance under the new conditions of assortative mating. Two preliminary points may be briefly noted to begin with.

(*a*) *As regards Environment.* It is here particularly important to distinguish between environmental effects that are random and those that may be described as systematic. With investigations of the type that concern us here we shall be confronted in the main with environmental effects which are almost exclusively random, and therefore likely to resist any endeavour to eliminate them, whether by an appropriate selection of the initial data or by some suitable statistical technique. Certainly in the case of bodily measurements we may reasonably assume that the investigator, in planning his research and selecting his sample, will avoid any cases in which environmental influences and genetic influences are liable to be markedly correlated. And in the case of psychological measurements we may hope to achieve the same purpose by a suitable choice of tests, though here, as we shall find, our success is less complete.

(*b*) *As regards Dominance.* In dealing with dominance we are compelled to recognize that it is no longer practicable to exclude all irrelevant correlations by deliberate planning or selection. Thus, with multifactorial inheritance, as we have already noted, whenever there is an appreciable correlation between observed measurements for the fathers and mothers, there must be, almost inevitably, a certain amount of correlation between similar phases of the different factors. Hence in computing the covariances, as in computing the variances, it will no longer be safe to deduce expressions for a single factor considered in isolation, and then simply add the results to secure corresponding expressions for the case of multifactor inheritance.

These two considerations entail the introduction of two new constants for environment and for dominance respectively.

The Multipliers for Environment and for Dominance. To simplify the discussion, let us formulate it in terms of some concrete characteristic such as height. As measured in actual practice, the height obtained for any given individual will include the effects both of environment and of dominance. Let us therefore imagine what the measurement would be in theory were it feasible to remove first the one and then the other of these two irrelevant influences. Let us write x_0 for the phenotypic height of the individual as actually measured; x_1 for the hypothetical measurement which we should presumably have obtained had his height developed under standard environmental conditions exactly the same for all; and x_2 for the hypothetical genotypic measurement we should have secured had we been able to assess it directly in terms of the linear scale defined above ($c-b$, c, $c+b$). And now let us build up the regression functions [1] which express the reciprocal relations between x_1 and x_0 and again between x_1 and x_2.

Think first of the bivariate frequency distribution for x_1 and x_0. Take any group or 'array' of persons whose height, measured under standard environmental conditions, is the same for each individual, x_1' (say). Then, in accordance with our preliminary assumptions, the actual environmental differences, being purely random, will in the long run tend to neutralize one another: consequently, the expected value for the mean phenotypic height of the individuals in the array selected will be $\bar{x}_0 = x_1'$. On the other hand, if we take an array of persons whose phenotypic height is the same for each individual, x_0' (say), then the expected value for their mean height as measured under standard environmental conditions would be $\bar{x}_1 = c_1 x_0'$, where c_1 (the regression coefficient) will be given by the ratio

$$\frac{\text{variance with environment uniform}}{\text{variance augmented by random environmental differences}} = \frac{V_g}{V_t}, \text{ or } h^2 \qquad (19)$$

in the notation borrowed above from Wright. Evidently, when we are dealing with a characteristic (like eye-colour) which is wholly unaffected by environment, or when we have so planned our inquiry that even random environmental differences are excluded, c_1 will be unity.

Now consider the bivariate frequency distribution for x_1 and x_2. Here, if we pick an array of individuals for each of whom $x_2 = x_2'$ (i.e., some value that is the same for all), we should expect their mean x_1 to be $\bar{x}_1 = x_2'$. On the other hand, if we take an array of individuals for each of whom $x_1 = x_1'$, we should expect their mean x_2 to be $\bar{x}_2 = c_2 x_1'$, where c_2 (as before) denotes the ratio between (i) the variance without the variations due to dominance and (ii) the variance with those deviations. Our previous formulae for the correlation between the hypothetical genotypes (eqn. 17, 18) were derived before we had taken into consideration the possible effects of environment on the coefficient for assortative mating. To allow for this we must now write

$$A = c_1 c_2 \mu. \qquad (20)$$

Evidently, when the effects of dominance are wholly negligible, $\epsilon^2 = 0$, and therefore $\tau^2 = \sigma^2$, and c_2 will be unity.

Application to Correlations. We have now to determine the net result of these influences on the correlations between relatives. But before we can do so, we have first to settle a question originally raised by Karl Pearson [44]: are we to assume that the phenomena of assortative mating arise in consequence of more or less conscious preferences for certain phenotypic tendencies (i.e., are they based on x_0), or do they arise through some less obvious causal process, e.g., selective fertility (in which case the primary correlation would be between the values for x_2 in husband and wife)? With characteristics of the kind with which we are here chiefly concerned (height and intelligence, for example), it seems safer to adopt the former assumption. In this case, if μ denotes the marital correlation for x_0 (the measurements actually obtained), the marital correlation for x_1 will be $c_1\mu$, and that for x_2 will be $c_1 c_2 \mu = A$.

The mean x_2 of a sibship will of course be equal to the mean x_2 of the parents. And for any array whose observed measurements are x_0', the mean $\bar{x}_1$ will be $c_1 x_0'$, and the mean $\bar{x}_2$ will be $c_1 c_2 x_0'$. For their mates, the mean x_0 will be $\mu x_0'$, the mean x_1 will be $c_1 \mu x_0'$, and the mean x_2 will be $c_1 c_2 \mu x_0 = A x_0'$. Hence the mean x_2 of the children will be $c_1 c_2 \frac{1+\mu}{2} x_0$.

Now the correlation between parents and children depends (as we have already shown) solely on the genotypic values x_2. It follows that the parental correlation will be

(i) $$r_{po} = c_1 c_2 \frac{1+\mu}{2}. \qquad (21)$$

[1] We use the term 'regression' in the broader sense adopted by Fisher [21, chap. v, sect. 25].

Again, since the mean x_2 of an array from the first generation (i.e., the children) is $c_1c_2\frac{1+\mu}{2}x_0'$, it follows that the mean x_2 of their mates, when they grow up and marry, will be $c_1c_2\frac{1+\mu}{2}Ax_0'$. Thus the correlation between the original parents and the second filial generation, i.e., the grandparental correlation, will be

(ii) $$r_{go} = c_1c_2\frac{1+\mu}{2}\cdot\frac{1+A}{2}. \tag{22}$$

Continuing in this way, we find that the correlation for the $(n+1)$th generation will be

(iii) $$c_1c_2\frac{1+\mu}{2}\left(\frac{1+A}{2}\right)^n. \tag{23}$$

This, as Fisher points out, yields *a* '*law of ancestral heredity*' *as deduced from a multifactorial mode of inheritance*. And, as we hope later on to show, it appears to fit the coefficients reported by most recent observers much better than the formulae derived by Galton or by Karl Pearson from somewhat different assumptions [44, pp. 475-481 and refs.].

In the same way we obtain for the fraternal correlation

(iv) $$r_{oo'} = \tfrac{1}{4}(c_1+c_1c_2+2c_1^2c_2^2\mu). \tag{24}$$

It will be observed that, if mating is random, $\mu = 0$; and in that case the parental correlation (eqn. (21)) reduces to $r_{po} = \frac{1}{2}c_1c_2$, or $\frac{1}{2}h^2g^2$ in the notation introduced by Wright: similarly the fraternal correlation (eqn. (23)) reduces to $r_{oo'} = \frac{1}{4}c_1(1+c_2)$ or $\frac{1}{4}h^2(g^2+1)$ in Wright's notation: (cf. eqn. (12) and (13) above).

Of the other correlations between collaterals two are of special interest to the psychologist, since they provide a convenient empirical test for hypotheses like those we are examining, namely, (v) the correlation between uncles (or aunts) and nephews (or nieces),

$$r_{un} = c_1c_2\left(\frac{1+A}{2}\right)^2+\frac{1}{8}c_1A(1-c_2), \tag{25}$$

(vi) the correlation between first cousins,

$$r_{kk'} = c_1c_2\left(\frac{1+A}{2}\right)^3+\frac{1}{16}c_1A^2(1-c_2). \tag{26}$$

The expressions thus obtained, it will be noted, indicate that the last two correlations differ but little from the correlations for grandparents and great-grandparents respectively.

Linked Genes and Multiple Allelomorphs. We shall not attempt to carry our theoretical investigations beyond this point. But perhaps a word or two is necessary about the effects of removing the two other important restrictions with which we started. In constructing our mathematical model we assumed that the genetic factors were inherited in complete independence, i.e., we have entirely ignored the consequences of linkage. And it is not difficult to show that linkage, if present, would certainly modify the formulae for the second degree statistics, such as variances and covariances [cf. Mather, 34, chap. v]. With sex-linkage we should expect the correlations for siblings and for cousins to differ when computed separately for male, female, and mixed pairs: our data do in fact show such differences; but with this we shall deal in a later paper. We also assumed that each genetic factor was capable of only two alternative forms. But the possibility of multiple allelomorphs cannot be altogether excluded: serological studies, for example, have demonstrated that in man the four major blood groups imply the existence of three or more genes, occurring at the same locus, and therefore all allelomorphic to each other; and in dealing with quantitative characters in somewhat humbler organisms several recent writers [1] have suggested as an alternative starting-point the assumption of a few genes each with many 'iso-alleles' instead of many genes each with only two allelomorphs. Fisher, however, has devoted special attention to the effects of polymorphic genes, and concludes that 'the introduction of multiple allelomorphism does not affect the simplicity of the final results'.

[1] Cf. E. C. R. Reeve and F. W. Robertson, 'Studies in Quantitative Inheritance I-IV'. *J. Genet.*, L, 1951, pp. 414-15 and later vols. If, however, as was suggested above, we leave the concrete nature of our 'factors' unspecified, the statistical deductions appear to apply equally to either version of the multifactor theory. See also K. Mather, 'The Genetical Units of Continuous Variation', *Proc. IXth Internat. Congr. Genetics*, 1954, pp. 106-138.

VII. VERIFICATION OF FORMULAE

Alternative Models. Most of the formulae set out above were originally put forward by Fisher as alternatives to those previously accepted by Karl Pearson and other members of the biometric school. Now many psychological writers (e.g., Lewis [29]) still seem to think in terms of much the same mathematical model as Karl Pearson : they tacitly assume complete dominance and equal proportions for the two pure types (dominant and recessive), and ignore the influence of assortative mating. On this basis, as Pearson himself showed, the theoretical correlations obtained are in sharp disagreement with the coefficients actually reported. It is therefore not surprising that later psychological critics who have accepted Pearson's view of heredity should have endorsed Pearson's conclusion, and summarily rejected the multifactorial version of the Mendelian theory.

Yule [66] modified Pearson's initial assumptions by postulating not complete dominance, but complete absence of dominance ; and this model, no doubt because of its extreme simplicity, has in the past been the one most frequently adopted for elementary expositions of the subject, particularly in discussing experimental work on animals and plants. It is, however, not to be wondered at if the psychological critic of today proceeds to argue that, although Yule's postulates may be adequate for certain characteristics in humbler organisms, there is as yet little evidence to justify their use as a basis for multifactorial theories in man.

Curiously enough hardly any of the numerous writers who have criticized the theory of innate differences in intelligence have considered the third form of the multifactorial hypothesis, namely, that which allows for varying degrees of dominance, although that appears to be the version now most widely accepted by geneticists themselves in discussing other forms of quantitative inheritance.

Our first task, therefore, must be to take these three alternative versions, and see which of them seems best suited to describe the transmission of quantitative characteristics in man. We shall begin by considering bodily characteristics, since here the initial measurements themselves are easier to secure and less likely to be called in question.

(i) *Stature.* Stature happens to be the quantitative variable first used to test theories of human inheritance (Galton, *Natural Inheritance*, 1889) ; it is also the human characteristic

TABLE IX. STATURE : THEORETICAL CORRELATIONS COMPARED WITH OBSERVED

Relatives	Theoretical			Observed
	Random Mating		Assortative Mating	
	Dominance Complete	Dominance Absent	Dominance Incomplete	
(*a*) Direct Line				
With Parent	·333	·500	[·507]	·507
With Grandparent	·166	·250	·301	·322
With Great-grandparent	·083	·125	·189	—
(*b*) Collaterals·				
Between Siblings	·416	·500	·536	·543
Between Uncles (or Aunts) and Nephews (or Nieces)	·166	·250	·301	·287
Between First Cousins	·083	·125	·181	·236

for which the most extensive data are available. We give, in the last column of Table IX, the correlations for stature actually observed in the case of relatives of various degrees of affinity. In this column the correlations between parent and offspring and between siblings are taken from Pearson and Lee [45], who measured well over 2,000 pairs. The other three coefficients are based on our own data, which, like those of Pearson and Lee, were collected from families of the University student class: here the numbers are much smaller (263, 615, and 584 pairs respectively).

The figures in columns 1, 2, and 3 are based on the hypotheses of Pearson, Yule, and Fisher respectively. Thus the values shown in the first two columns are derived from the first two assumptions, namely, random mating with (1) complete dominance and (2) complete absence of dominance. The figures in column 1 are identical with those deduced by Pearson himself [46]. It will be seen that, in both the first and the second columns, the coefficients (with one exception) proceed by geometrical progression, each being one half of its predecessor. Those for the direct line in the second column agree with Galton's ' Law of Ancestral Heredity ' (based on data obtained for human stature and for coat colour in Bassett hounds), namely, that the correlation with the nth direct ancestor is $(\frac{1}{2})^n$.

To compute the theoretical values in the third column we require three supplementary parameters: c_1 (expressing the effects of environment), c_2 (expressing the degree of dominance), and A (expressing the effects of assortative mating). The figures given by Pearson and Lee yield a marital correlation of $\mu = \cdot 2804$, and a parental correlation of $r_{po} = \cdot 5066$. Substituting for r_{po} and μ in eqn. (21), we obtain $c_1c_2 = \cdot 7913$, and $A = c_1c_2\mu = \cdot 2219$. Using these two results we can at once compute the theoretical values for relatives in a direct line. To compute the values for siblings and other collaterals we require the value of c_1 as well as c_1c_2. However, recent work on twins [1] indicates that, unless the range of environmental differences is extreme, the stature of adults is but little affected by post-natal conditions. Hence we are not likely to go far wrong if we provisionally take $c_1 = 1$, particularly as it occurs by itself in only one of the three terms. The last three theoretical coefficients in column 3 have therefore been computed on this basis.

There can be little question which of the three alternative versions of the multifactorial hypothesis best fits the data observed. The coefficients derived by Pearson, as he himself was the first to emphasize, evince a wide divergence from those actually observed: they are in fact invariably too low. If we assume that dominance is completely absent, we obtain a better fit; but even here most of the discrepancies are statistically significant. When we take into account the effects both of assortative mating and of incomplete dominance, we obtain values so close that we may justifiably claim that, for stature at any rate, the multifactorial theory, as Fisher has formulated it, is here amply confirmed.

By way of a further check let us use the value actually obtained for the fraternal correlation to re-determine c_1. Substituting in eqn. (24) the values actually observed, we obtain $c_1 = 1 \cdot 03$; the true value of course cannot exceed unity: but this difference falls well within the margin allowed by the sampling error. Had we assumed that the marital correlation was due primarily to a preference or selection (perhaps half unconscious) which affects, not so much the phenotypic characteristic x_1 but the essential genotype x_0, then the estimate for c_1 would sink to about 0·88. Since it might plausibly be maintained that the marital correlation is probably due to both types of selection, perhaps the most reasonable estimate would lie somewhere between the two, which would give $c_1 = \cdot 95$ (approx.).

Fisher extends the verification to cover the data given by Pearson and Lee for other physical measurements, such as span and arm-length; and, so far as the available observations go, the empirical figures are everywhere in close conformity with the theoretical.

(ii) *Intelligence.* Let us now inquire how far the theoretical values calculated by the methods we have advocated fit the observed correlations for assessments of intelligence. The sources for the latter, the procedures employed, and the results obtained have already been described in previous publications (Burt, 1955, and refs.). The observed correlations are entered in the last column of Table X.

The assessments for fathers and mothers yield a marital correlation of ·386 (a figure which we are inclined to think a little too high). To compute theoretical values for the parental and fraternal correlations, we also need estimates for the two parameters c_1 and c_2. These we might assess in round figures from the conclusions already reached with physical measurements. Accordingly, let us provisionally put c_1 (the parameter indicating the effects of non-heritable components) at 0·95; and let us suppose that the different genes vary between complete dominance and complete absence of dominance, and so put c_2 midway at 0·75. With these rough estimates we obtain the two theoretical values, shown in brackets, in Table X, col. 2. They are in fairly close agreement with the correlations actually observed; but in both cases the theoretical estimates are rather too high. As we might expect, owing to unreliability and the like, the effect of the non-heritable components is evidently greater with mental measurements than with physical, so that c_1 ought to be lowered. However, with samples of this size the differences are not significant.

[1] The correlation between identical twins reared in different environments is 0·969, which is not significantly different from that for identical twins reared together, 0·932 ([39]; cf. [7]).

TABLE X. INTELLIGENCE: THEORETICAL CORRELATIONS COMPARED WITH OBSERVED

Relatives	Number	Theoretical: Assortative Mating; Dominance Incomplete	Observed
(*a*) Direct Line			
With Parents	963	[·495]	·489
With Grandparents	321	·311	·335
With Great-grandparents	—	·198	—
(*b*) Collaterals			
Between Siblings	987	[·514]	·507
Between Uncles (or Aunts) and Nephews (or Nieces)	375	·309	·354
Between First Cousins	552	·215	·289

To estimate the remaining correlations, let us now substitute *empirical* estimates for c_1 and c_2 deduced directly from the paternal and fraternal correlations actually observed. We then obtain the theoretical values shown in Table X, col. 2. This obviously furnishes a more satisfactory check. However, for these remoter relatives the observed correlations are based on still smaller numbers, and therefore on a selection of families that may possibly be not quite comparable with those in the larger sample: and the assessments for the adults are perhaps less trustworthy than those for the school children. Nevertheless, the theoretical values are again in close agreement with the observed. The only large discrepancy occurs in the case of first cousins. But here too none of the differences is statistically significant.

So far as our data go, therefore, the validity of the multifactorial hypothesis seems fully confirmed.

VIII. ANALYSIS OF VARIANCE

Procedure. With plants and animals successful attempts have recently been made to estimate an ' index of heritability ' (or similar coefficient) for various quantitative characteristics of agricultural or economic value—the yield of crops, the body-weight of sheep or pigs, the laying capacity of poultry, the yearly milk records of cows, and so forth. Thus, for weight of beef cattle at 15 months the index is estimated at 0·9 ; for butterfat production in dairy cattle, it is only 0·5. Such figures have proved to be of practical utility as a guide to selective breeding and in the estimation of future productivity. It is tempting therefore to suggest that somewhat similar indices might be usefully computed for human traits, particularly those of educational, industrial, or social importance.

The obvious procedure is to partition the total variance for the trait in question into the main genetic and non-genetic components (cf. Mather [34], Wright [50, 64]). The statistical model we have described assumes that we can distinguish at least four additive components : (i) the main [1] fixable genetic component (or ' essential genotypes ', as Fisher calls them), G ; (ii) a non-fixable genetic component (chiefly the effects of dominance), D ; (iii) a component due to assortative mating, M ; (iv) an environmental or non-genetic component (which Fisher assumes will be, with the data chosen for his analysis, a random component), E ; and possibly (v) a joint residual term. The equations reached above should enable us to deduce the requisite formulae for each of the four principal components.

Let us, as before, denote the total variance by V_t and the variance due to genetic constitution only by V_g. Since c_1 expresses the ratio between the two (eqn. 19), we have at once $V_g = c_1 V_t$. The variance of the non-genetic component will therefore be

$$k_e E = (1-c_1)V_t. \qquad (27)$$

[1] The qualifying term ' main ' is inserted here because this first component does not include that portion of the ' fixable genetic component ' which enters into the third component (assortative mating). In the brief side headings to Tables XI and XII this qualification is of course to be understood.

From eqn. (10) and (16), we obtain

$$V_g = \sigma^2 + \frac{A}{1-A}\tau^2 = \tau^2 + \epsilon^2 + \frac{A}{1-A}\tau^2 = \frac{\tau^2 + (1-A)\epsilon^2}{1-A}.$$

But by eqn. (10) and (20) we have

$$c_2 = \frac{\tau^2}{\tau^2 + (1-A)\epsilon^2}.$$

Hence,
$$\frac{\tau^2}{V_g} = \frac{\tau^2(1-A)}{\tau^2 + (1-A)\epsilon^2} = c_2(1-A),$$

that is,
$$k_g G = \tau^2 = c_2(1-A)V_g = c_1 c_2 (1-A)V_t. \tag{28}$$

Again,
$$k_d D = \epsilon^2 = \epsilon^2 \frac{(1-A)}{\tau^2 + (1-A)\epsilon^2} \cdot \frac{\tau^2 + (1-A)\epsilon^2}{(1-A)}$$
$$= (1-c_2)V_g = c_1(1-c_2)V_t. \tag{29}$$

Finally,
$$k_m M = \frac{A}{1-A}\tau^2 = \frac{A}{1-A} c_2(1-A)V_g = c_2 A V_g = c_1 c_2 A V_t. \tag{30}$$

These formulae, however, assume that we already know the values for the three essential parameters, c_1, c_2, and A. As before, we may assign them approximate values on the basis of our general antecedent knowledge of the traits and populations concerned, though this is perhaps a little precarious. Alternatively we may derive empirical values for them, provided observed figures are available for three of the relevant correlations—the parental, the fraternal, and the marital: in that case the requisite values can be computed from the following expressions, derived from eqn. (20), (21), and (22):

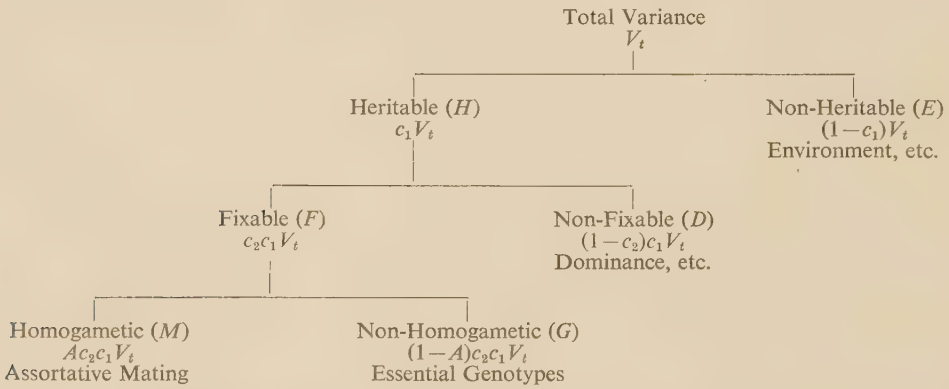
$$c_1 c_2 = \frac{2r_{po}}{1+\mu}, \tag{31}$$

$$A = c_1 c_2 \mu, \tag{32}$$

and
$$c_1 = 4r_{oo'} - c_1 c_2 (1+2A). \tag{33}$$

Since the right-hand members of the four equations 27-30 together add up to V_t, we can in theory reconstruct the total variance from these four contributions.[1] The analysis, it will be seen, is based on a succession of ' subdivided factors ' each of which is ' bipolar '. The resulting scheme may be set out as follows (Fig. 1).

FIGURE 1. ANALYSIS OF VARIANCE INTO ITS PRINCIPAL COMPONENTS



[1] Fisher does not himself tabulate the algebraic values; and the tabular analysis for stature that he offers by way of incidental illustration does not explicitly include any estimated figure for the effects of environment, i.e., it assumes that $c_1 = 1$. In the text, however, he observes that such effects might perhaps account for something up to 5 per cent. of the total variance, though he thinks the amount ' very unlikely ' to be quite as large as this.

(i) *Stature.* To illustrate the general procedure we shall first calculate the contributory variances for stature, on the basis of the data collected by Pearson and Lee [45]. For the parental and fraternal correlations we have used the values given by Fisher: these were obtained by averaging the correlations given by Pearson and Lee for the two sexes separately, since the sex differences were insignificant. We thus have for our starting point $r_{po} = \cdot 5066$, $r_{oo'} = \cdot 5433$, and $\mu = \cdot 2804$, from which we can calculate $A = \cdot 2219$. For c_1 probably the best estimate is given by the mean value of ·9553, obtained by averaging two alternative estimates computed by Fisher from slightly different assumptions. This implies that nearly 95 per cent. of the variance exhibited by a normal sample of the population is due to genetic constitution and barely 5 per cent. to non-genetic causes (Table XI).

We arrive at much the same figures by an alternative route if we consider the fraternal correlation, ·5433. Regarded as a ratio of two variances, such a figure indicates that at least 54 per cent. of the total variance is directly attributable to ancestry, namely, that portion of the total variance as a result of which siblings from the same family *resemble* one another. But from the

TABLE XI. ANALYSIS OF VARIANCE: STATURE

Source	Formula	Percentage of Total Variance
Genetic Component: Fixable .	$c_1c_2(1-A)V_t$	61·57
Genetic Component: Non-fixable .	$c_1(1-c_2)V_t$	16·40
Assortative mating . . .	$c_1c_2AV_t$	17·56
Environment	$(1-c_1)V_t$	4·47
Total	V_t	100·00

theoretical formula deduced from the Mendelian hypothesis (eqn. 13) we infer that much of the *difference* between siblings is also due to genetic constitution, namely, that portion which depends on the genetic factors in respect of which the parents are heterozygous. According to our equation this is measured by $\frac{1}{2}\tau^2+\frac{3}{4}\epsilon^2$, and here therefore takes the value $\cdot 3078+\cdot 1230 = \cdot 4308$. On this basis, which is perhaps a little less exact, we should conclude that $(54+43) = 97$ per cent. of the total variance is due to genetic constitution, leaving only 3 per cent. to be accounted for by non-genetic factors.

These figures of course hold good only for the particular type of population sampled, namely University students and their relatives. Data obtained from London school children indicate that in boroughs where (at the time of our earlier surveys) poverty and malnutrition were rife, the non-genetic variance might amount to nearly 20 per cent. Here, however, our purpose is not to inquire how far an unfavourable environment can impair the inherited constitution of an individual child, but merely to determine how widely the individual's constitution may vary under reasonably uniform and approximately normal conditions and how seriously our current methods of assessing such variations are liable to be affected by irrelevant influences.

(ii) *Intelligence.* We now come to our main problem. To assess intelligence as we have defined the term, it will be unwise to rely exclusively on formal tests of the usual type, particularly when they consist largely of verbal group tests and have been applied on a single occasion only. Provided the person tested is in good health at the time of the examination, has a strong motive for doing his best, and is in no way handicapped by special disability, emotional disturbance, or a lack of the elementary knowledge which the problems presuppose, then such a test may no doubt furnish a tolerably accurate assessment. But this cannot be taken on trust; and the only way to be sure that no distorting influences have affected the results is to submit the marks to some competent observer who has enjoyed a first-hand knowledge of the testees. With children this will usually be the school teacher; and whenever discrepancies appear between the teacher's verdict and that of the test, the child must be re-examined individually, preferably with tests of a non-verbal type.

Such precautions were regularly adopted by the psychologist working in the London County Council schools, since one of his main duties was concerned with the ascertainment of possible cases of certifiable deficiency or of scholarship ability which had been overlooked or questioned.

The interview, the use of non-verbal tests, and the information available about the child's home circumstances usually made it practicable to allow for the influence of an exceptionally favourable or unfavourable cultural environment. Whether the effects of infantile ailments or irregular growth could always be allowed for is perhaps more questionable. In any case, when test-results have been systematically checked and adjusted by these means, the reliability of the final assessments, and their correlation with the pupils' subsequent achievements, prove to be far higher than those of a single intelligence test, whether group or individual.

Even so, however, the reliability coefficients still indicate that the measurements thus obtained contain a far larger element of error than those for physical characteristics, like height or span. And it is no longer possible to assume that all the systematic effects of environment—physical, physiological, or social—have been successfully eliminated. This means that, in dealing with psychological assessments, we shall have to introduce at least two further components into our statistical model to take account of these additional disturbances.

The Estimation of Non-heritable Variation. In the general population there is undoubtedly an appreciable correlation between environmental conditions and what are ordinarily classed as inheritable differences. For this reason indeed the environmentalist contends that, for all we know, the real cause of such differences may be simply the environmental conditions themselves. ' Wealth and ability ', says Dr. Heim, ' tend to go together ' ; and the wealthier parent is not only himself better educated, but is also in a position to send his child to a good preparatory and public school, which in turn will improve the child's ability: ' given better food and more books at home, the poorer children would match, if not outstrip, their more fortunate fellows on the tests '.[1] But of course it would be equally possible to turn the argument round, and maintain that the hereditary differences were the primary cause of the differences in the environmental background; and perhaps still more plausible to hold that there was a reciprocal interaction between the two.

There are two obvious ways in which we may endeavour to forestall these objections. We may supplement the material obtained from unselected samples by data obtained from selected groups in which either (*a*) the relevant environment or (*b*) the genetic components are approximately constant.

Pearson's data for stature, as we have seen, were obtained from a fairly homogeneous population, where the economic and material conditions in the homes were much the same for all. And, in pursuance of the same principle, the psychologist may collect his mental measurements from orphanages or residential homes where the children (often illegitimate, and therefore widely differing in heredity) are brought up under practically the same conditions from early infancy. Results so obtained have been fully summarized elsewhere.[2] But in studies planned on Mendelian lines the alternative approach is far more instructive. In designing experiments on plants and animals, it is customary to start, as Mendel did, by selecting inbred ' pure lines ' for the initial cross. Under these conditions any minor variations in the measurements either of the true-breeding parents or of their immediate offspring (F_1) must be attributable exclusively to environment. Such variations will thus serve to indicate how far the phenotypic trait (or our method of assessing it) is affected by non-hereditary fluctuations. With human beings this course is scarcely feasible. But fortunately

[1] This argument has been much to the fore in recent discussions on equality of educational opportunity. Those who rely on it are perhaps a little too ready to assume (i) that the education received at a preparatory or public school is in fact far more efficient than that received at a good secondary school of the kind attended by the scholarship winner or his present-day equivalent, and (ii) that there is an appreciable ' transfer of training ' whereby the superior instruction given at such schools effects an all-round improvement in the general intelligence of the pupil : at the same time they overlook the possibility that a superiority in inheritable intelligence may not only itself be the main reason for the superior home-environment, but also, when transmitted to the child, enable him in his turn to make more profitable use of his educational opportunities. But we have no intention of thrashing out these broader questions here : we merely wish to point out that the final decision must depend on many other lines of evidence than those which are relevant to the narrower issue discussed in this paper.

[2] Cf. Burt (1943), pp. 90 f. and refs. Dr. Heim quotes Burt's article in the passage cited in the text, but attributes his conclusion to ' conservative-minded prejudice ' in favour of ' the supremacy of heredity and the rightness of the existing order '. She seems almost to picture him as a psychological Pangloss, always maintaining that *tout est au mieux*. She omits, however, to indicate where precisely the figures supporting his conclusion are wrong: indeed, few of her readers, we imagine, would guess from her comments that the conclusion itself was based on a systematic analysis of actual data.

with them an equivalent source of evidence is available: instead of selecting monozygotic strains we can select monozygotic twins.

Evidence from Twins. As Galton was the first to realize, and later investigators have so often pointed out, the occurrence in man [1] of 'identical' or one-egg twins 'makes it possible to show more clearly than in other animals the relative influence of differences in genotype and in environment respectively' ([11], p. 349). And in previous inquiries we ourselves have relied largely on data of this type. However, several of our critics have argued that, since twins are born at the same time and are, as a rule, inseparable companions, the similarity in their pre-natal and post-natal environment must be much greater than with ordinary siblings, and might thus be responsible for the larger correlations [24, 25]. As commonly stated, this contention is largely met by comparing (*a*) monozygotic twins with dizygotic twins and (*b*) monozygotic twins brought up in the same home with those brought up in separate homes.[2] Yet, even when brought up in separate homes, it is said [31], such children must still have been subjected to identical intra-uterine conditions, equally favourable for both in some cases, and equally unfavourable in others. This latter assumption, however, is quite unwarranted. Because twins have developed within the same uterus during the same period of time, it does not for a moment follow that the pre-natal conditions must have been the same for both. To begin with, the position of the foetuses in the uterus, and the time (and often the mode) of delivery, will differ. Moreover, the immediate cause of uniovular twinning seems in most instances to be some temporary setback at the very outset of gestation; and this is likely to affect the different foetuses in different degrees. Before splitting, the two portions of the developing embryo generally develop at different rates; and, when splitting supervenes at a rather advanced stage, the twin that arises from the less developed half tends to be both smaller and weaker. If the splitting occurs after the embryo has begun to exhibit bilateral asymmetry, then one twin may be, in certain anatomical details, the mirror-image of the other: in 'Siamese' and other conjoined twins, where the splitting occurs so late as to remain incomplete, such reversals seem the rule rather than the exception. In our own series we noted that, with one-egg twins, opposite handedness (one left-handed, the other right-), reversed patterns in the fingerprints, contrary directions in the crown whorl of the hair, were nearly four times as frequent as in two-egg twins. In two instances one showed *situs inversus viscerum* (reversed position of heart, liver, etc.). It was among the pairs which furnished evidence for late splitting that we have commonly encountered the largest discrepancies in the estimates for intelligence: as a rule, the child delivered second turns out to be the feebler of the two: (indeed, the second is the one most frequently still-born). Our conclusion

[1] Uniovular twins are occasionally reported in the case of sheep and swine; but on the whole the production of offspring from split ova seems much rarer among other animals than in man. The example most fully studied is that of the nine-banded armadillo of Texas, which regularly produces 'identical' quadruplets from the splitting of a single fertilized egg: and the detailed examination of the process in that creature has done much to clarify our knowledge of the whole problem: (cf. J. T. Patterson, 'Polyembryonic Development in *Tatusia novemcincta*', *J. Morphol.*, XXIV, 1913, pp. 120-145).

[2] Referring to the results of Newman *et al.* [39], Dr. Heim declares that, 'when monozygotic twins are separated,... the differences between them are as great as those between unseparated dizygotic twins' [24, p. 130]. This, however, is not borne out by the correlations printed by Newman *et al.* themselves. Throughout her discussion of their material she fails to distinguish between test performance and innate capacity—a mistake which the original investigators carefully avoid. The most striking difference—that between 'Helen' who had a Stanford-Binet I.Q. of 116 and 'Gladys' who scored only 92—a difference quoted by several of our critics, is plainly attributable to the fact that Helen 'graduated from a good college and taught in school for 12 years', while Gladys was reared 'in an isolated part of the Rockies where there were no schools'. Anyone who was familiar with the researches by Hugh Gordon, Frances Gaw, and others on canal-boat and rural children would realize that the Stanford-Binet scale cannot possibly be accepted as providing a fair estimate of *innate* ability with those brought up in such different environments—least of all with adults.

Newman himself points out that each of the four pairs showing a wide disparity in test-performances had been subjected to very different types of education, and notes incidentally that 'even with a good education a poorly endowed person cannot reach the level of a potentially able but poorly educated person'. He concludes that 'in each of these cases we must assume that the twin with the lower I.Q. had an *inherited* capacity to reach the rating of the other' [40, p. 146].

May we take this opportunity of appealing for further information about 'identical' twins brought up in different homes? Such cases are far commoner than is generally supposed. Frequently the mother herself feels unable to look after two children at once. But the most usual ground for the separation is the fact that the relatives or friends who come to the rescue on the mother's death are unwilling to accept responsibility for more than one. Owing however to the popular prejudice against separating twins, there is a natural desire to keep such decisions secret—a secrecy which we should of course scrupulously respect.

therefore is that intra-uterine conditions are, if anything, more likely to diminish the resemblances between twins than to increase them, but that their effect on mental capacity (save in the more extreme cases) is comparatively slight.

Children of Consanguineous Marriages. The objection just discussed might be partly overcome by comparing siblings born at different times to parents who are closely related—e.g., the offspring of first cousins. In London, and particularly in Council schools, children of consanguineous unions seem rarer than elsewhere. Among the medical profession, for example, Pearson found that 4·5 per cent. of the marriages were between first cousins; and others have reported figures nearly as high among 'the upper middle classes' and 'members of the aristocracy'. On the other hand, in a survey covering all the pupils at Council schools in a typical London borough, Burt [2] could discover only 32 such cases—less than 0·9 per cent. In later surveys and by other means we have since been able to locate 255 pairs of such children, making 287 in all, mostly from better class families. On assessing their intelligence by the methods described above, we find a correlation of ·623 as compared with ·507 for ordinary siblings. The increase, as we might expect, is far smaller than in the case of identical twins; yet it is statistically significant.[1] Now, if similarity of environment, either before or after birth, played an appreciable part in producing correlations between the assessments for siblings, it should exert just the same influence on those who are not of consanguineous parentage as on those who are: (the fact that the consanguineous pairs are drawn from a narrower range of social conditions could only tend to lower the correlation, not to raise it). On the other hand, if differences in intelligence are mainly of genetic origin, then some such increase in correlation is precisely what we should expect; and it is hard to see what other explanation could be assigned. Meanwhile, it seems reasonable to assume that the correlation obtained for identical twins who have lived apart almost all their lives will yield the most direct estimate for the proportional contribution of genetic factors to the variance of whatever measurements we are making.

Analysis of Variance for Twins. Our own data for twins include test-results and carefully checked assessments both for binovular and for monovular pairs, and the latter include twins brought up together by their own parents and a smaller number separated in early infancy. There are various ways in which such information may be used to throw light on the relative influence of heredity and environment.

1. Keeping first to the larger groups, namely, the pairs brought up together, we may regard each pair as a family of two, and thus estimate, for either group, the amount of variance *within* families. With the binovular twins, these variations will result partly from differences in genetic constitution and partly from differences in the environmental conditions or from inaccuracy in the assessments. With uniovular twins there can be no genetic difference between a pair from the same family: hence any observable difference will be due solely to the latter types of cause. Now, if r denotes the *intraclass* correlation for a given group, then $r = (V_b - V_w)/(V_b + V_w)$, where V_b and V_w denote the variances between and within families respectively (Burt [4], p. 275). Hence, writing $V_t = V_b + V_w$ for the total variance, we have $V_w/V_t = \frac{1}{2}(1-r)$. Let us affix the subscripts 1 and 2 to denote the values obtained for the uniovular and binovular twins respectively; and let us assume (i) that the contribution of environment to the variance within families is the same for both groups of twins, and (ii) that the total variance for each group is also the same. Then the proportional amount of the intra-familial variance contributed by 'nature' or heredity (H), as contrasted with 'nurture' or environment (E), will be

$$\frac{H}{H+E} = \frac{V_{w2} - V_{w1}}{V_{w2}} = \frac{r_1 - r_2}{1 - r_2}.$$

With the correlations obtained from our data, this gives, for the unadjusted test marks,

$$\frac{\cdot 904 - \cdot 542}{1 \cdot 000 - \cdot 542} = 79 \cdot 0 \text{ per cent., and for the adjusted assessments, } \frac{\cdot 925 - \cdot 551}{1 \cdot 000 - \cdot 551} = 83 \cdot 3 \text{ per cent.}$$

[1] The difference becomes still more striking if the influence of preferential mating is first eliminated from the correlation for ordinary siblings (either by statistical methods or by a suitable selection of families). The coefficient obtained for the consanguineous siblings, like that for first cousins themselves, is appreciably higher than would be deduced from the formulae cited above. It is possible that in the offspring of such matches the effect of unifactorial recessives helps to augment the correlation.

It is to be hoped that other investigators will watch for data of this kind. Pedigrees of families springing from consanguineous unions have been assiduously used for the study of qualitative abnormalities, particularly those associated with recessive genes; they might with equal advantage be used in studying quantitative variables, like intelligence.

For Stature the corresponding figure is 91·8 per cent., for Reading 31·7 per cent., for Spelling 36·2 per cent., and for Arithmetic 45·4 per cent.[1]

The assumptions adopted are perhaps not entirely free from criticism; and the procedure itself is hardly suited to our own particular needs. There is no explicit allowance for unreliability, and no attempt to distinguish systematic from random environmental influences; nor can we assume that a proportion based on the intra-familial variance of twins is the same as the total variance of the general school population. Nevertheless, so far as they go, the proportionate values suggest two definite conclusions: first (and here our own figures are strongly reinforced by those of Newman, Holzinger, and Freeman), the contribution of heredity to the variance of ordinary group tests of intelligence is very much higher than that for educational assessments, yet not so high as for physical characteristics; secondly, its contribution to the variance of the adjusted assessments is distinctly higher than that of the raw and unadjusted marks for such tests taken just as they stand.

2. Since, in addition to the data just analysed, our material includes results for uniovular pairs brought up in different environments, and, in a fairly large and random sample, yields two independent assessments for the same individuals, it should be possible to make at least some approximate estimate (*a*) for the systematic effects of environmental influences as distinct from the random, and (*b*) for the unreliability involved in all such measurements. We can thus amplify the conclusions reached with the foregoing procedure. By way of a provisional determination, the following simplified method has been suggested by Miss Howard, who has also undertaken the detailed calculations. The reliability coefficient may itself be regarded as a ratio of two variances; and accordingly we may write, in the customary notation, $r_{xx'} = V_g/(V_g+V_e)$, where V_g and V_e denote the variances of the common and error factors respectively. This at once yields an estimate for the amount of the total variance attributable to 'error', namely, $V_e/(V_g+V_e) = 1-r_{xx'}$. Moreover, if the foregoing arguments be accepted, we may agree that, when identical twins are brought up in separate environments from the earliest years, then the similarities between them will owe little or nothing to similar or systematic variations in the environment, and may therefore be attributed solely to their genetic constitution. Thus the correlation between such pairs may be used to estimate the proportional contributions of genetic and non-genetic influences respectively. The variance due to the former, we shall assume, can now be subdivided into three contributory components in the proportions indicated by Fisher's formulae; and we shall also use the same formulae to indicate the proportion attributable to 'random' environmental influences.[2]

We have applied this procedure both (*a*) to the adjusted assessments of intelligence, on which we have mainly relied both here and in our previous discussions, and (*b*) by way of comparison, to the unadjusted marks obtained by a single application of the tests. The initial correlations are: (*a*) with adjusted assessments, for twins reared apart, ·8756; for parent and child, ·4887; siblings, ·5069; father and mother, ·3875; reliability, ·9476; (*b*) with group tests before the marks were corrected or adjusted, for twins reared apart, ·7706; for parent and child, ·5288; siblings, ·5342; father and mother, ·4526; reliability, ·9357. The final results are set out in Table XII. The figures in the first and third columns give an analysis of variance along the lines already adopted for stature. Those in the

[1] Except that our argument is expressed in terms of intraclass correlations and of variances between and within families, this procedure is practically the same as that originally suggested by Holzinger [27]. In the studies on twins that he carried out with Newman and Freeman, the coefficients obtained were, for the Otis group tests, 80 per cent., for Stature, 81 per cent., for Finger ridges, 90 per cent., for Spelling, 53 per cent., and for Arithmetic, 12 per cent. [39].

[2] There are also other methods of analysis which deserve consideration. Shuttleworth [52, cf. 58], for example, made free use of partial correlation and supplemented it by Sewall Wright's device of path coefficients. But this procedure does not lend itself so readily to an analysis along Mendelian lines. Wright himself has since outlined a method, not unlike Fisher's, based on an analysis of variance [64].

The method described above and used in Table XII has been adopted here partly because in a first attempt it seemed wiser to follow a procedure which would be at once simple and readily intelligible, and partly because it leaves the main proportions directly comparable with those obtained for physical characteristics by Fisher's formulae. It is, however, admittedly a makeshift. The main difficulty springs from the large number of variables involved and the complexities that arise when we attempt to allow for their possible interactions and intercorrelations. Many of these could, we believe, be rendered negligible by improved methods of planning the inquiry and selecting the necessary data. In the earlier stages it seems undoubtedly best to concentrate solely on the more important and better attested source of data, even at the cost of simplification and a possible risk of increased error. The use of simplified models inevitably invites attack from armchair critics. Nevertheless, in other branches of biometric research, such a policy has already been advocated and ably defended by Mather and others [13].

second and fourth columns show the more detailed analysis obtained with the fuller procedure just described.

(*a*) *Adjusted Assessments.* Let us glance first at the proportions based on the raw correlations (column 1), since these are directly comparable with the figures for height. The proportion due to the 'fixable' genetic component (or, as Fisher expresses it, the 'essential genotypes') is here but little over 50 per cent., whereas in the case of stature it exceeded 60 per cent. On the other hand, the amount attributed to the 'non-fixable genetic component' (i.e., deviations due to dominance) is appreciably larger, possibly because this component here includes other intrinsic sources of irregularity. The proportions attributed to assortative mating and to the 'random' effects of environment

TABLE XII. ANALYSIS OF VARIANCE: INTELLIGENCE

Source	Adjusted Assessments		Unadjusted Group Tests	
	Using Raw Correlations	With Allowance for Environment and Unreliability	Using Raw Correlations	With Allowance for Environment and Unreliability
Genetic component: fixable . .	51·35	47·92	48·82	40·51
Genetic component: non-fixable .	23·28	21·73	20·07	16·65
Assortative mating .	19·19	17·91	23·99	19·90
Environment: random	6·18	5·77	7·12	5·91
Environment: systematic . .	—	1·43	—	10·60
Unreliability . .	—	5·24	—	6·43
Total . .	100·00	100·00	100·00	100·00

(which in cols. 1 and 3 include all effects uncorrelated with the genetic tendencies) are, as we might expect, also increased, though not to so large an extent.

It is, however, the figures in the second column that answer, or attempt to answer, the questions that chiefly interest the practical psychologist. With adjusted assessments for intelligence, obtained in the way we have described, about 11 per cent. of the total variance appears to be attributable either to the random effects of the environment or to unreliability, and less than 2 per cent. to the systematic effects of the environment, i.e., to effects working in the same direction as the genetic deviations: (on the whole, this is what we ourselves should expect, since the methods of assessment used were deliberately designed to exclude any such differences). That leaves about 87 per cent. to be ascribed to genetic constitution.

(*b*) *Unadjusted Marks.* Let us now contrast these figures with the parallel set obtained from the uncorrected marks for the tests (last column). Here nearly 23 per cent. of the variance is apparently due to non-genetic influences, i.e., to environment or to unreliability; and over 10 per cent. now seems directly attributable to the systematic effects of the environment.

The Parallelogram Formula. If the method of analysis we have here proposed is reasonably accurate, it should be possible with its aid to estimate the extent to which the relevant environmental differences are correlated with innate or genetic differences between the children and with the tests and the adjusted assessments we have used to estimate such differences. This may enable us, not only to check the accuracy of the analysis, but also to determine which particular aspects of the environment have the greatest influence on the children's performances in our tests.

The method or model best suited to this purpose is one which treats the variables concerned as vectors combining in accordance with the parallelogram law. It thus provides an alternative and equally instructive way of representing the whole problem.

Consider first the figures obtained for the crude or ' unadjusted ' marks furnished by the group tests. Taking the total variance as 100·0 (as in the last column of Table XII) let us begin by deducting the amount attributable to ' unreliability ', since that is irrelevant to our immediate problem. Then the remainder, namely, 93·57 (which we may term the ' reduced total ') can be subdivided into three main portions : (i) that due to heredity, 77·06 ; (ii) that due to the independent or ' random ' effects of environment, 5·91 ; and (iii) a residual portion, 10·60, which we have attributed to the ' systematic ' effects of environment. Evidently we could plot the first two hypothetical components as if they were two forces at right angles to each other ; the lengths of the lines representing them will be proportional to their standard deviations, that is, to $\sqrt{(77{\cdot}06)} = 8{\cdot}78$ and $\sqrt{(5{\cdot}91)} = 2{\cdot}43$ respectively. We now assume a supplementary environmental influence, acting in the same direction as the component of heredity, and thus reinforcing the influence of the latter. In the diagram the joint effect of these two environmental tendencies will be represented by a single line inclined at an angle to the vertical axis (since that merely represents the independent or ' random ' effects). The direction and amount of the resultant of all these forces, represented by a line making an angle with each of the two axes and having a length equal to the standard deviation of the test results, can now be determined by completing the parallelogram, and drawing its diagonal. (The student will find it instructive to make an actual diagram for himself.)

In the present instance the length of the diagonal is already given, namely, $\sqrt{(93{\cdot}57)} = 9{\cdot}67$. From this, together with the three data specified above, we can at once deduce the length of the supplementary component, and the size of the several angles involved. After a little elementary calculation we find that the line representing the joint environmental effects must be inclined at an angle of 76° 29′ with the axis representing heredity, and therefore at an angle of 61° 56′ with the line representing the marks for the tests. The former implies a correlation of cos 76° 29′ (i.e., ·234) between the joint environmental effects and heredity ; the latter implies a correlation of cos 61°56 (i.e., ·471) between the joint environmental effects and the tests. In reconstructing the reduced total variance (σ_t^2 say), therefore, we are implicitly adopting the familiar formula

$$\sigma_t^2 = \sigma_h^2 + \sigma_e^2 + 2r_{he}\sigma_h\sigma_e, \tag{34}$$

that is, with the values just specified,

$$93{\cdot}57 = (8{\cdot}78)^2 + (2{\cdot}50)^2 + 2 \times 0{\cdot}234 \times 8{\cdot}78 \times 2{\cdot}50.$$

If we turn to the figures for the adjusted assessments (second column of Table XIII), an analogous computation leads to the following equation :

$$94{\cdot}76 = (9{\cdot}36)^2 + (2{\cdot}40)^2 + 2 \times 0{\cdot}032 \times 9{\cdot}36 \times 2{\cdot}40.$$

In this case the correlation between heredity and the joint environmental effects is ·032, that is, virtually zero : while the correlation between the assessments and the joint environmental effects is ·277.

How far do the calculated correlations agree with the results observed? Unfortunately it was impossible to procure ratings for environmental conditions [1] in the case of the entire group. However, ratings of socio-economic status are available for 253 of the families ; and this gives a fairly adequate indication of the way cultural conditions vary in what appears

[1] In a paper on ' Environmental Effects in Quantitative Inheritance ' [50, pp. 81-102] Dr. Bernard Woolf not only questions the validity of the whole statistical procedure, but also the conception of environment which it appears to imply. Fisher and his followers, so he maintains, use the term environment ' as an abstraction, denuded of any real reference to the life conditions of the child in its home, school, and play '. After criticizing both Burt and Thomson he observes : ' one wonders what meaning, if any, these authors attach to the term environment '. If any of our readers raise the same question, may we refer them to the two chapters on ' Environmental Conditions ' in Burt's *Young Delinquent* (pp. 62-206) and the corresponding chapters in *The Backward Child* (pp. 89 f.)? These chapters discuss in considerable detail the nature of the relevant environmental conditions, ' in the home, in the school, and in the child's leisure hours at play ', and describe the various ways in which such conditions apparently affect the mental development of the child. Their direct influence on tests of various types is discussed in earlier publications (e.g., *Mental and Scholastic Tests*, pp. 102 f., ' Influence of Social Status ').

In considering the alternative hypothesis (preferred by many earlier critics as well as Dr. Woolf)—namely, that differences in ' intelligence ' might be due wholly or mainly to environmental conditions—we argued (i) that differences in school conditions were only likely to affect badly chosen or badly constructed tests, (ii) that home environment might affect verbal tests, but not non-verbal, (iii) that ill-health in early child-hood (pre-natal and post-natal) might affect all estimates of intelligence, but that its influence was comparatively rare and nearly always slight. The fact that pupils suffering from all these handicaps may nevertheless attain scholarship level is alone sufficient to refute the explanation championed by Dr. Woolf.

to be a representative sample. For this smaller group the correlation between the ratings and the crude test-results is ·453 and between the ratings and the adjusted assessments ·315 (Burt, 1943, p. 84). It will be seen that the two observed coefficients agree reasonably well with the estimated correlations for ' joint environmental conditions ' deduced by the method just described. We thus have a further confirmation of the validity of the theoretical analysis to which we have been led by our modified version of Fisher's procedure.[1]

In conclusion, we should like to insist on the very limited nature of the problem we have tried to solve. Neither here nor elsewhere have we attempted (as so many critics have apparently supposed) to reach any over-all statement about the relative contributions of heredity and environment to mental efficiency or ' intelligence ' as manifested in ordinary everyday life or among all classes and conditions of men. Our inquiry has been concerned solely with an analysis of assessments obtained by the procedures we have described and of test-results supplied by ' group tests ' of the kind most psychologists use: the question we have sought to answer is not, how much does ' heredity ' contribute to individual differences in intelligent behaviour generally, but simply, how much does it contribute to the assessments obtained by this or that particular method of estimating genotypic differences. Moreover, our analysis holds only for a population of the particular type we have sampled, brought up in an environment of a certain definite and restricted character. We readily admit that there still remain several obvious defects in the procedure itself, e.g., the absence of satisfactory tests for statistical significance. However, most of the objections raised by our previous critics appear to turn, not on incidental imperfections of this kind, but on a general misunderstanding either of the problem or of the hypothesis. With these broader criticisms we hope to deal in a later paper. Meanwhile, we should be the first to agree that, pending more extensive and better planned inquiries carried out along more or less similar lines by other investigators, our own conclusions, even with the limitations we have stressed, can be accepted as no more than provisional.

IX. SUMMARY AND CONCLUSIONS

We have endeavoured to formulate in more precise detail the multifactorial theory of the inheritance of intelligence tentatively put forward by one of us in several previous publications. In the case of human beings the verification of such a hypothesis is bound to rest almost entirely on statistical evidence. Accordingly, in the hope of providing at least a provisional method of analysis, we have outlined the procedures originally devised by Fisher for dealing with the inheritance of bodily measurements and developed by later geneticists like Mather and Sewall Wright; at the same time we have ventured to introduce one or two minor modifications of our own, partly to meet the peculiarities of psychological data and partly as a result of more recent criticism and research. The formulae so reached for the distribution of intelligence, both univariate and bivariate, and for the correlations between relatives of varying degrees, yield theoretical values which are in close accord with the empirical figures calculated from results of actual surveys ; whereas those deduced from other interpretations fail to account for most of the coefficients so obtained. To that extent therefore the hypothesis seems vindicated.

Previous attempts to estimate what proportion of the total variance is attributable to genetic constitution (or ' heredity ' as it is commonly expressed) appear usually to have been based solely on the extent of the *resemblances* between relatives and overlook the fact that *differences* among members of the same families may also be due to genetic influences. On analysing the variance obtained with carefully checked assessments of ' intelli-

[1] To avoid misunderstanding we should say that we do not regard the manner in which the test results are apportioned between ' heredity ' and ' environment ' as in any way final. The appearance of a positive correlation between the child's cultural environment and his success in the group tests might be due, not so much to the possibility that the more favourable environment may have helped the child to answer the tests, but rather to the fact that the intelligence of the parents (which led them to supply the child with an intellectual home background) has itself been transmitted biologically to the child, and so indirectly enabled him to tackle the test-items more intelligently. If so, the proportion of the variance assigned to the influence of heredity has been somewhat underestimated: it includes only the direct influence, not the indirect.

BIBL. LONDIN. UNIV.

gence', we find that with the method we have adopted about 12 per cent. is apparently attributable to unreliability and to irrelevant environmental influences, and the rest to genetic constitution (including the effects of dominance and assortative mating). With the scores for group tests taken just as they stand, we find that as much as 23 per cent. is apparently due to unreliability and environment, and only about 77 per cent. to genetic constitution.

These conclusions, if confirmed, raise a number of important practical questions still unsolved: how far, for example, and by what methods, can we allow for or correct the inaccuracies in current methods of testing and assessing the innate abilities of children, whether for educational, clinical, or vocational purposes? Is there any way of assessing the innate intelligence of adults, e.g., by reference to the assessments made in childhood or to the family history? What, if any, are the social and political implications of the theory here developed? On all these questions numerous conflicting pronouncements, based for the most part on *a priori* preconceptions and unverified deductions, have been put forward in the past by hereditarians and by environmentalists respectively; but hardly ever have the issues been impartially investigated in and for themselves. Here surely lies one of the most urgent problems that the statistical psychologist can be called upon to study.

REFERENCES

1. Burt, C. (1912). 'The inheritance of mental characteristics.' *Eugen. Rev.*, IV, 168-200.
2. Burt, C. (1914-30). *Reports of the Psychologist in the Education Officer's Department.* London County Council.
3. Burt, C. (1935). *The Subnormal Mind.* London: Oxford Univ. Press.
4. Burt, C. (1940). *The Factors of the Mind.* London: Univ. London Press.
5. Burt, C. (1943). 'Ability and income.' *Brit. J. Educ. Psychol.*, XIII, 83-98.
6. Burt, C. (1953. *The Contributions of Psychology to Social Problems.* London: Oxford Univ. Press.
7. Burt, C. (1955). 'Evidence for the concept of intelligence.' *Brit. J. Educ. Psychol.*, XXV, 158-177.
8. Carmichael, L. ed. (1946). *Manual of Child Psychology.* New York: Wiley.
9. Clark, P. J. (1956). 'The heritability of certain anthropometric characters as ascertained from measurements of twins.' *Amer. J. Human Genetics*, VIII, 46-63.
10. Conrad, H. S. and Jones, H. E. (1940). 'Environmental and genetic implications of parent-child and sibling correlations.' *Yearbook Nat. Soc. Stud. Educ.*, XXXIX (ii), 97-141.
11. Darlington, C. D. and Mather, K. (1949). *The Elements of Genetics.* London: Allen & Unwin.
12. David, F. N. (1949). *Probability Theory for Statistical Methods.* Cambridge Univ. Press.
13. Dunn, L. C,. ed. (1941). *Genetics in the Twentieth Century.* New York: Macmillan.
14. East, E. M. and Emerson, R. A. (1913). 'The inheritance of quantitative characters in maize.' *Res. Bull. Agric. Sta. Nebraska*, II.
15. East, E. M. and Jones, D. F. (1919). *Inbreeding and Outbreeding.* Philadelphia: Lippincott.
16. 'Espinasse, P. G. (1942). 'The polygenic concept.' *Nature*, CXLII, 732-734.
17. Fisher, R. A. (1918). 'The correlation between relatives on the supposition of Mendelian inheritance.' *Trans. Roy. Soc. Edin.*, LII, 399-433.
18. Fisher, R. A. (1919). 'The causes of human variability.' *Eugen. Rev.*, X, 213-220.
19. Fisher, R. A. (1930). *The Genetical Theory of Natural Selection.* Oxford: Clarendon Press.
20. Fisher, R. A., Immer, F. R. and Tedin, O. (1932). 'The genetical interpretation of statistics of the third degree in the study of quantitative inheritance.' *Genetics*, VII, 107-124.
21. Fisher, R. A. (1934). *Statistical Methods for Research Workers.* Edinburgh: Oliver & Boyd.
22. Fisher, R. A. (1949). *The Theory of Inbreeding.* Edinburgh: Oliver & Boyd.
23. Hardy, G. H. (1908). 'Mendelian proportions in a mixed population.' *Science*, XXVIII, 49-50.
24. Heim, A. W. (1954). *The Appraisal of Intelligence.* London: Methuen.
25. Hogben, L. (1929). 'The relative effects of nature and nurture on twin differences.' *J. Educ. Psychol.*, XX.
26. Hogben, L. (1933). *Nature and Nurture.* London: Williams & Norgate.
27. Holzinger, H. J. (1929). 'The relative effect of nature and nurture influences on twin differences.' *J. Educ. Psychol.*, XX, 243-256.
28. Johannsen, W. (1909). *Elements der exakten Erblichkeitslehre.* Jena: Fischer.
29. Lewis, D. G. (1956). 'The normal distribution of intelligence: a critique.' *Brit. J. Psychol.* (In press.)
30. Lush, J. L. (1943). *Animal Breeding Plans.* Iowa State College Press.

31. Maddox, H. (1956). 'Nature-nurture balance sheets.' *Brit. J. Educ. Psychol.* (In press.)
32. Mather, K. (1943). 'Polygenic inheritance and natural selection.' *Biol. Rev.*, XVIII, 32-64.
33. Mather, K. (1948). 'The genetical theory of continuous variation.' *Proc. VIIIth Internat. Congr. Genetics*, 376-401.
34. Mather, K. (1949). *Biometrical Genetics: The Study of Continuous Variations.* London: Methuen.
35. Mendel, G. (1866). 'Versuche über Pflanzen-Hybriden.' *Verhandlungen des Naturforschenden Vereines in Brünn*, IV, 3-43.
36. National Society for the Study of Education (1928). *Twenty-Seventh Yearbook.* Bloomington, Ill.: Public School Publishing Co.
37. Neel, J. V., Kodani, M., Brewer, R., and Anderson, R. C. (1949). 'The incidence of consanguineous marriages in Japan, with remarks on the estimation of gene frequencies.' *Am. J. Human Genetics*, I, 156-178.
38. Neel, J. V. and Schull, W. J. (1954). *Human Heredity.* Chicago: Univ. Chicago Press.
39. Newman, H. H., Freeman, F. N., and Holzinger, K. J. (1937). *Twins: A Study of Heredity and Environment.* Chicago: Univ. Chicago Press.
40. Newman, H. H. (1942). *Twins and Supertwins.* London: Hutchinson.
41. Nilsson-Ehle, H. (1908). 'Einige Ergebnisse von Kreuzungen von Hafer und Weizen.' *Botan. Notiser Lund*, 287-302.
42. Nilsson-Ehle, H. (1909-11). 'Kreuzungsuntersuchungen am Hafer und Weizen.' *Lunds Univ. Arsskrift*, II, v, 1-122, vii, 1-84.
43. Panse, V. G. (1940). 'A statistical study of quantitative inheritance.' *Ann. Eugen.*, X, 76-105.
44. Pearson, K. (1900). *Grammar of Science.* London: A. & C. Black.
45. Pearson, K. and Lee, Alice (1903). 'On the laws of inheritance in man. I. The inheritance of physical characters.' *Biometrika*, II, 357-462.
46. Pearson, K. (1904). 'On a generalized theory of alternative inheritance with special reference to Mendel's laws.' *Phil. Trans.*, CCIII, 53-87.
47. Pearson, K. (1909). I. 'The theory of ancestral correlations in heredity.' II. 'On the ancestral gametic correlations of a Mendelian population mating at random.' *Proc. Roy. Soc.*,B, LXXXI, 219-225.
48. Penrose, L. S. (1935). 'Genetic linkage in human graded characters.' *Ann. Eugen.*, VIII, 233-238.
49. Punnett, R. C. (1905). *Mendelism.* Cambridge: Macmillan & Bowes.
50. Reeve, E. C. R., and Waddington, C. H., eds. (1952). *Quantitative Inheritance.* London: H.M. Stationery Office.
51. Roberts, J. A. Fraser (1940). *An Introduction to Medical Genetics.* Oxford: Oxford Univ. Press.
52. Shuttleworth, F. K. (1935). 'The contributions of nature and nurture to individual differences in intelligence.' *J. Educ. Psychol.*, XXVI, 655-681.
53. Smith, H. H. (1944). 'Recent studies on inheritance of quantitative characters in plants.' *Botan. Rev.*, X, 349-382.
54. Spuhler, J. N. (1948). 'On the number of genes in man.' *Science*, CVIII, 279-280.
55. Srb, A. M. and Owen, R. D. (1952). *General Genetics.* San Francisco: Freeman.
56. Stone, C. P. (1947). 'Methodological resources for the experimental study of innate behaviour as related to environmental facts.' *Psychol. Rev.*, LIV, 342-347.
57. 'Student' (1934). 'A calculation of the minimum number of genes in Winter's selection experiment.' *Ann. Eugen.*, VI, 77-82.
58. Vernon, P. E. (1950). 'Psychological studies of the mental quality of the population.' *Brit. J. Educ. Psychol.*, XX, 35-42.
59. Weinberg, W. (1908). 'Über den Nachweis der Vererbung beim Menschen.' *J. Verein f. vaterl. Naturk. in Würtemburg*, LXIV, 368-382.
60. Weinberg, W. (1909). 'Über Vererbungsgesetze beim Menschen.' *Zeit. f. indukt. Abstamm. u Vererbungslehre*, I, 277-330.
61. Weinberg, W. (1910). 'Die Anlage zum Mehrlingsgeburt beim Menschen.' *Arch. Rass. u. Ges. Biol.*, VII, 35-49, 169-173.
62. Wingfield, A. H. (1928). *Twins and Orphans: The Inheritance of Intelligence.* London: Dent.
63. Wright, Sewall (1934). 'An analysis of variability in number of digits in an inbred strain of guinea pigs.' *Genetics*, XIX, 506-536.
64. Wright, Sewall (1935). 'The analysis of variance and the correlations between relatives with respect to deviations from an optimum.' *J. Genetics*, XXX, 243-256.
65. Yule, G. U. (1902). 'Mendel's laws and their probable relation to intra-racial heredity.' *New Phytologist*, I, 193-207, 222-237.
66. Yule, G. U. (1906). *Report IIIrd Internat. Congr. Genetics*, 140-142.
67. Yule, G. U. and Kendall, M. G. (1937). *Introduction to the Theory of Statistics.* London: Griffin.

APPENDIX. THE CALCULATION OF MENDELIAN FREQUENCIES BY MATRIX METHODS

The construction of bivariate frequency-tables for relatives of various degrees, and of the resulting correlations, can be greatly simplified by adopting a matrix notation for the tables and operations involved. The following outline is appended to illustrate the possibility in an elementary form. By borrowing the more technical but less familiar conventions of group-theory, more efficient formulations could readily be devised, but would be scarcely intelligible to the ordinary reader. We require symbols for—

I. MATRICES (denoted by descriptive initials, small for vectors, i.e., single rows or columns, and capital for rectangular matrices). Five main classes of matrices may be distinguished.

1. *Metric or Scale Matrices.* E.g., $s' = [i, j, k]$.
2. *Type Matrices.* E.g., $g' = [\mathbf{AA}, \mathbf{Aa}, \mathbf{aa}] = [D, H, R]$ for genotypes.
3. *Frequency Matrices.* E.g., $f' = [p^2, 2pq, q^2]$.

For these, see Table III, cols. 1, col. 3 and 4, and col. 6, respectively.

4. *Combined Matrices.* E.g., $c' = [p\mathbf{A}, q\mathbf{a}]$.
5. *Product Matrices.* Of these the most important will be matrices representing (i) *Recombinations*, e.g., $G_2 = gg'$ (genotypes of the 2nd filial generation, cf. Table I), (ii) *Mating Systems*, e.g., $M_f = ff'$ (cf. Table IV), and (iii) *Correlation or Contingency Tables*, e.g., $F_{1f} = ff'$ (frequencies for the 1st filial generation, cf. Table VII).
6. *Numerical Matrices*, chiefly matrices of numerical constants which often recur in varying contexts, e.g. the probabilities that result from random pairing, e.g., $N_1 = [N_{\mathbf{AA}}, N_{\mathbf{Aa}}, N_{\mathbf{aa}}]$ (the constants in Table VI, col. 3 written as a 3×9 matrix), and N_2, the constants in the squares of these expressions (cols. 4-9) written as a 9×9 matrix.

II. OPERATIONS (denoted by algebraic signs, superscripts, or Greek letters). In addition to the usual matrix operations of (i) *Adding* $(+)$, (ii) *Multiplying* $(\times)$, and what may be termed (iii) *External Squaring*, i.e., post-multiplying an entire matrix by its transpose $(^2)$, and (iv) *Internal Squaring or Powering*, i.e., raising the elements of the matrix to a higher power $(\mu^2$, etc.), it is convenient to designate as distinct operations the results (v) of forming a mating system, i.e., *Pairing* (π), (vi) of forming a contingency table, i.e., *Co-relating* (χ). Since paper-work is limited to two-dimensional schemes, it is also useful to have symbols for processes of rearrangement (which have no analogue in ordinary algebra), viz., in addition to the familiar process of (vii) *Transposing* $(')$, those of (viii) *Aligning*, i.e., rewriting a square matrix as a row or column (λ), (ix) *summing*, i.e., pre-multiplying by the vector $[1, 1, \ldots, 1] = \sigma'$, and (x) *Recombining*, i.e., classifying recombinations (ρ).

The frequencies for mating systems may also be dealt with by using 'generating functions': e.g., Table IV and the first 2 cols. of Table VI can be obtained by expanding $(p\mathbf{A} + q\mathbf{a}]^4 = p^4\mathbf{AA} \times \mathbf{AA} + 2p^3q\mathbf{AA} \times \mathbf{Aa} + 2p^3qAa \times aa + \ldots$ etc. (For matrix work it is simpler to keep such tables as Table VI expanded to 9 rather than 6 terms.) Similarly cols. 4-9 (expanded to 9) can be obtained by treating the rows of col. 3 as generating functions and squaring. The values in Table VII will then be obtained by premultiplying the values in cols. 4-9 by those in col. 1 (transposed to a row vector).

If the reader will attempt to express in symbolic form such processes as finding the correlation between parent and child $(f'f.N_1)$, siblings, $\lambda(f'f).N_2$, or between cousins $X(ff')\sigma' . \pi[\sigma' . (ff')\rho(ff')]$, he will find the result increases clarity and speed of computation. By introducing further conventions even simpler expressions may be derived (as the reader will quickly discover for himself; and by starting with double vectors (one for each sex) the method may readily be generalized to cover problems or sex-linkage.

INDEX TO VOLUME IX

INDEX OF AUTHORS

Authors	*Titles*	*Page*
Audley, R. J., and Jonckheere, A. R.	Stochastic Processes and Learning Behaviour	87–94
Baggaley, A., and Cattell, R. B.	A Comparison of Exact and Approximate Linear Function Estimates of Oblique Factor Scores	83–86
Banks, C.	Journal Committee	67
Barton, D. E., and David, F. N.	Spearman's ' Rho ' and the Matching Problem	69–73
Burt, C.	Obituary: Professor L. L. Thurstone	1–4
Burt, C.	Weighted Summation applied to Paired Comparisons	63–65
Burt, C., and Foley, E.	The Statistical Analysis of the Learning Process	49–62
Burt, C., and Howard, M.	The Multifactorial Theory of Inheritance and its Application to Intelligence	95–131
Curr, W., and Gourlay, N.	Differences between Testers in Terman-Merrill Testing	75–81
McQuitty, L. L.	Agreement Analysis: Classifying Persons by Predominant Patterns of Responses	5–16
Peel, E. A.	The International Colloquium on Factor Analysis	66–67
Rimoldi, H. J. A.	Prediction of Scale Values for Combined Stimuli	29–40
Slater, P.	Weighting Responses to Items in Attitude Scales	41–48
Watson, H. E.	Agreement Analysis: A Note on Professor McQuitty's Article	17–20
Woodbury, M. A., and Lord, F. M.	The Most Reliable Composite with a Specified True Score	21–28

INDEX OF SUBJECTS

Titles	*Authors*	*Page*
Agreement Analysis: A Note on Professor McQuitty's Article	Watson, H. E.	17–20
Agreement Analysis: Classifying Persons by Predominant Patterns of Responses	McQuitty, L. L.	5–16
Attitude Scales, Weighting Responses to Items	Slater, P.	41–48
Composite, The Most Reliable, with a Specified True Score	Woodbury, M. A., and Lord, F. M.	21–28
Factor Analysis, International Colloquium on	Peel, E. A.	66–67
Inheritance, Multifactorial Theory of, and its Application to Intelligence	Burt, C., and Howard, M.	95–131
Journal Committee	Banks, C.	67
Learning Behaviour, Stochastic Process and	Audley, R. J., and Jonckheere, A. R.	87–97
Learning Process, Statistical Analysis of	Burt, C., and Foley, E.	49–62
Oblique Factor Scores, Comparison of Exact and Approximate Linear Function Estimates of	Baggaley, A., and Cattell, R. B.	83–86
Paired Comparisons, Weighted Summation Applied to	Burt, C.	63–65
Scale Values for Combined Stimuli, Prediction of	Rimoldi, H. J. A.	29–40
Spearman's ' Rho ' and the Matching Problem	Barton, D. E., and David, F. N.	69–73
Terman-Merrill Testing, Differences between Testers in	Curr, W., and Gourlay, N.	75–81
Thurstone, Professor L. L., Obituary	Burt, C.	1–4

BOOK REVIEWED

Author	*Title*	*Page*
Kendall, M. G.	Rank Correlation Methods	68